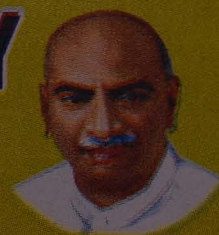




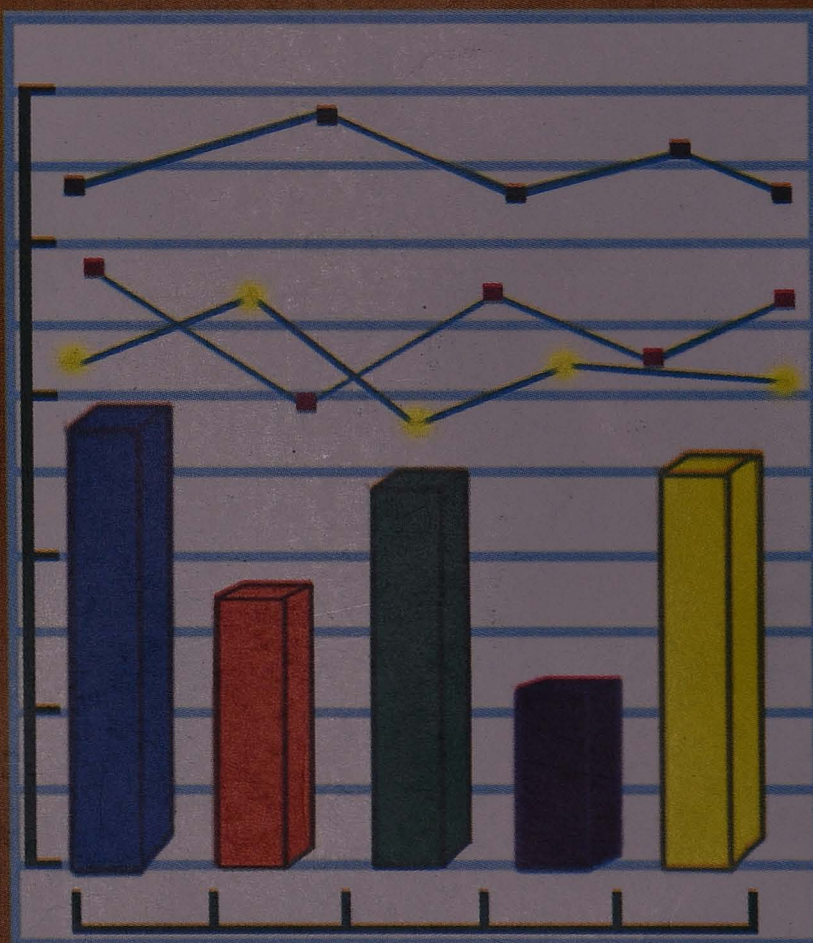
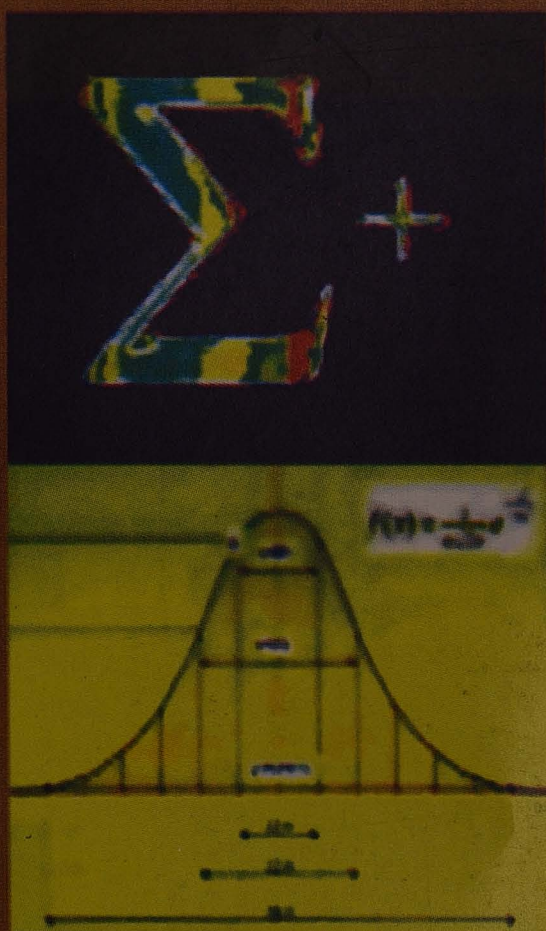
MADURAI KAMARAJ UNIVERSITY

(University with Potential for Excellence)



B.A.(Economics) Second Year

QUANTITATIVE TECHNIQUES VOLUME - I



DISTANCE EDUCATION

(Recognised by D.E.C.)

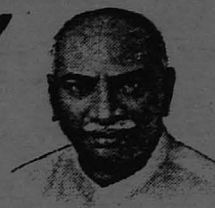
www.mkudde.org

S 50



MADURAI KAMARAJ UNIVERSITY

(University with Potential for Excellence)



485

B.A. (ECONOMICS)

Second Year

Paper - 1

QUANTITATIVE TECHNIQUES

Vol. I

Units 1 to 3

DISTANCE EDUCATION

(Recognised by D.E.C.)

www.mkudde.org

S 50

© All copyright privileges are reserved



MADURAI KAMARAJ UNIVERSITY

(University with Potential for Excellence)



482

B.A. (ECONOMICS)

Second Year

Paper - I

QUANTITATIVE TECHNIQUES

Vol. I

Units 1 to 3

DISTANCE EDUCATION

(Recognised by D.E.C.)

www.mkude.org

250

Printed at SAI GRAPHS April 2011- 500 copies

© All copyright privileges are reserved

B.A. (Economics) QUANTITATIVE TECHNIQUES

Second year

Dear student,

With our best wishes we welcome you to the Second Year B.A. (Economics) through distance mode. Quantitative Techniques is one of the three subjects which you have to study under Part III this year. You are expected to take an examination on this subject at the end of this academic year.

The Syllabus prescribed by the University for this subject is enclosed in this study material. For the sake of convenience, we have divided the whole syllabus into 10 Units and the elaborate scheme of these Units is also enclosed under the heading 'Contents'. All the 10 Units are presented in such a way that you will be able to learn the subject yourself without any external guidance.

As far as the subject, 'Quantitative Techniques' is concerned, you cannot learn the subject overnight. You have to study regularly and workout more and more problems on each topic. Therefore, as soon as you receive this book, start studying the Units one by one and try to work out all the problems given at the end of each Unit. If you work steadily, you can hope to get 100% easily in this subject.

Though we have taken our best efforts to present the subject matter in a lucid manner, you may find it difficult to understand certain concepts in some Units or even the whole unit. In order to help you to overcome such difficulties, we organise contact seminar programmes for you. The Annual Schedule of the contact seminar programme is sent to you individually and also it is made available in our website www.mkudde.org. You are free to attend the contact seminar at any centre. As you are receiving the study materials along with your Admission Card (provided you pay your Tuition fees in full), you should start studying them before you attend the contact seminar. Then only, you will be able to get full benefit from the contact seminar.

Expecting you to meet in the contact seminar, we wish you Best of Luck in all your endeavours.

**Faculty Members,
Department of Economics,
Distance Education,
Madurai Kamaraj University.**

UNIVERSITY SYLLABUS

PAPER - I : QUANTITATIVE TECHNIQUES

UNIT1:Functions and Equations:Variables, Functions and Equations-Types. Solving Quadratic Equation in one Variable and Simultaneous Linear Equations in Two Variables-Straight Line Equation - Slope of the Line - Homogeneous Function-Linear Homogeneous Function.

UNIT2:Calculus:Differentiation - Differentiation of Simple Functions- x^n , e^x , $\log x$ - First and Second Order Differential Coefficients - Maxima and Minima-Applications of Functions and Derivatives - Elasticities, Total, Marginal and Average Cost and Revenues.

UNIT3:Matrices:Matrices and Determinants-Types of Matrices-Value of Determinant -Inverse of a Matrix - Cramer's Rule.

UNIT4:Statistics:Meaning, Definition, Functions, Importance and Limitations of Statistics - Techniques of Data Collection - Sampling versus Census - Sampling Techniques - Primary and Secondary Data - Methods of Collection.

UNIT5:Averages:Mean, Median, Mode, Geometric Mean and Harmonic Mean.

UNIT6:Measures of Dispersion:Range, Mean Deviation, Quartile Deviation, Standard Deviation, Coefficient of Variation, Skewness and Kurtosis.

UNIT7:Correlation:Meaning and Types of Correlation - Measurement of Correlation-Scatter Diagram - Karl Pearson's Coefficient of Correlation - Spearman's Rank Correlation Coefficient.

UNIT8:Regression Analysis: Simple Linear Regression - Regression Equations - Properties of Regression Lines & Regression Coefficients - Uses of Correlation and Regression Analysis.

UNIT9:Index Numbers:Meaning, Definition and Types of Index Numbers-Construction of Wholesale Price Index Numbers - Laspeyre, Paasche and Fisher's Index Numbers-Tests for Ideal Index Number - Uses and Limitations of Index Numbers.

UNIT10:Time Series:Meaning, Definition and components of Time Series - Methods of Measurement of Trend - Moving Average Method and Method of Least Squares.

Books for Study

- 1. Introduction of Mathematical Methods - Bose, D. - Himalaya Publishing House.
- 2. Statistics-Pillai, R.S.N.&Bagavathi-S.Chand&Co.
- 3. Fundamentals of Applied Statistics-Gupta, S.P. and V.K. Kapoor-Sultan Chand & Sons.

References

- 1. Business Mathematics-D.C.Sancheti & V.K. Kapoor - Sultan Chand & Sons
- 2. Statistical Methods of Economic Analysis - Nagar, A.S. & Sharma, P.D. - S.Chand & Co.

CONTENTS

Volume I

Unit No.	TITLE	Page No.
1.	Nature and Scope of Statistics and Statistical Data Collection Procedures	1
2.	Averages	88
3.	Measures of Dispersion, Skewness and Kurtosis	206

Volume II

Unit No.	TITLE
4.	Correlation
5.	Regression
6.	Index Numbers
7.	Time Series
8.	Functions and Equations
9.	Matrices
10.	Differential Calculus
	About University Examination
	University Examination Question Papers

INTRODUCTION TO THE SUBJECT

'QUANTITATIVE TECHNIQUES'

There is virtually no branch of study which does not find the application of Mathematics and Statistics at present. Economics is a subject which makes wide application of both mathematics and statistics. For proper understanding, analysis and testing of various economic theories, application of various mathematical and statistical tools has become unavoidable. A study of various mathematical methods and statistical techniques put together is now-a-days called 'Quantitative Techniques' and it has been included as one among the three courses of study in your Second Year B.A. Degree Program.

The Syllabus prescribed by the University with reference to this course, 'Quantitative Techniques' consists of 10 units. The first three units consist of some of the basic mathematical techniques and their applications in Economics. The remaining seven units consist of the descriptive statistical tools widely used in Economics. We present the study material related to this course in Ten Units as detailed in the Contents.

In Unit-1, a brief introduction to the subject followed by a description about the nature, scope and importance of Statistics is given first. Then, the popular tabular presentation of data namely, Frequency Distribution and various concepts related to it are explained. In Unit-2, meaning and definition of average in general and of five important averages namely, mean, median, mode, geometric mean and harmonic mean are given. The computational procedures in respect of these averages are also explained in this Unit-2.

In Unit-3, meaning and need for various measures of Dispersion and their computational procedures are explained. Besides these, various measures of Skewness and Kurtosis are also explained in this Unit.

When two variables are given, to find out whether there is any relationship existing between the two variables, the statistical technique namely, 'Correlation' is useful and it is explained in Unit-4.

Given the value of one variable, say, advertisement expenditure of a firm, the corresponding value of a related variable, say, sales volume may be obtained with the help of the statistical technique known as 'Regression' and it is explained in Unit-5.

You may frequently come across the news items namely, 'steep rise in prices', 'inflation rate goes up' in India, in the daily news papers. How do we calculate the price rise? Index Numbers are the tools used for this purpose. Meaning, types and construction of such index numbers are explained in Unit-6.

How do or at what rate, the population of a country, production in various sectors of the economy etc. change or increase over the years? It is found out with the help of 'Time Series Analysis' which is explained in Unit-7.

Some mathematical tools like functions, equations, matrices, differential coefficients are widely used in Economics. All such tools are explained in Units-8, 9 and 10.

NATURE AND SCOPE OF STATISTICS AND STATISTICAL DATA COLLECTION PROCEDURES

Introduction :

The word statistics refers to the 'subject statistics' as well as to the 'numerical data'. Therefore, various people defined statistics in various ways. The important definitions alone we take into consideration and they are explained in this Unit-1. We use the word statistics to mean the subject which deals with the procedures to collect numerical data and the ways in which the collected data may be processed, analysed and interpreted. The data collection procedures and presentation of data in the form of Frequency Distributions are also explained in this Unit-1.

Unit Objectives :

After studying this Unit, you would be able to understand

- (i) the meaning, definition and main divisions of statistics
- (ii) the functions of statistics
- (iii) the importance of statistics in various fields especially with reference to Economics
- (iv) Limitations of statistics
- (v) various sources and methods of collecting statistical data
- (vi) the meaning and construction of frequency distribution.

Unit Structure :

1. Meaning, Definition and Divisions of Statistics
2. Planning and Conducting a Statistical Enquiry
3. Collection of Data or Sources of Data
4. Sampling
5. Frequency Distribution
6. Answers to the Check Your Progress Questions
7. Model questions for guidance :

1. Meaning, Definition and Divisions of Statistics

1.1 Meaning of the word Statistics:

The term STATISTICS is used in two senses, viz., (1) in a narrow sense and (2) in a wide sense. In the narrow sense, the word statistics denotes some numerical data. That is, it can refer to facts which can be put into a numerical

form, as in the phrase “unemployment statistics”, “statistics of industrial accidents in India”. This is the meaning the man in the street gives to the word statistics. It is to be noted that the word statistics used in the sense of numerical data is a plural noun.

In the wide sense, the word statistics refers to the statistical principles and methods which have been developed for handling numerical data. Statistical methods or statistics have a very wide range. They range from the most elementary descriptive devices like comparison, which may be understood by anyone, to the highly complicated mathematical produces which are comprehended by only the most expert theoreticians. This is the meaning in which we use the term statistics in our lessons. It is to be noted that the word statistics used in the sense of statistical principles and methods is a singular noun. Again, when we use the term statistics in the sense of statistical method it is a part of Applied Mathematics.

1.2 Subject-Matter of Statistics:

Statistics is concerned with the collection, presentation, analysis and interpretation of data which are measurable in numerical terms. It is to be noted that the facts which are dealt with in Statistics, must be capable of numerical expression. Crowden, Cowden and Klein have given the following illustration in their book ‘Applied General Statistics’, to emphasise this point. With the information that dwellings are build of brick, stone wood and other materials we cannot employ statistical methods. Instead, if we are given information regarding the number or proportion of dwellings constructed of each type of material, we have numerical data suitable for statistical analysis.

We saw above that statistics is concerned with the collection, presentation, analysis and interpretation of numerical data. Let us briefly examine each of these four procedures.

1.3 Four steps in a Statistical Analysis

1) Collection :

In any statistical investigation, collection of useful data is the first step. Data must be collected in a systematic manner. Data may be obtained form existing published or unpublished sources or by first hand collection.

2) Presentation:

Once collected, data must be assembled into a useful form. This process is the statistical presentation of data. Usually the data are arranged in tables or represented by graphic device.

Check your Progress

1. State the narrow meaning of the word 'Statistics'.
2. What is the subject matter of statistics.

3) Analysis:

By analysis of data we mean the study of the behaviour of the data. The basic characteristics of the data are studied with the help of statistical tools such as average, correlation, association etc.

4) Interpretation:

The final step in any statistical investigation is the interpretation of the data. Interpretation consists in deriving conclusions from the analysis of the data. One who interprets the data must do it bearing in mind the limitations of the original material.

1.4 Definition of Statistics as Numerical Statement of Facts

Various definitions of statistics have been given by different authors. Some of them have defined statistics as numerical statements of facts and others as statistical methods. We have examined below a few definitions of statistics.

Statistics definition as numerical statements of facts.

(i) Webster's definition of statistics is as follows:

"Statistics are the classified facts representing the conditions of people in a state... specially those facts which can be stated in numbers or in any tabular or classified arrangement".

This definition of Webster is too narrow, For, it restricts the scope of statistics to those facts and figures which related to the conditions of the people in a state. Again, this definition does not suit modern conditions. For, in the modern world, facts and figures are collected for studying all aspects of human activity and physical and biological phenomena also.

(ii) Prof. Horace Secrist has given a comprehensive definition of statistics, His definition of statistics is as follows:

"By statistics we mean aggregate of facts affected to a marked extent by multiplicity of causes numerically expressed, enumerated or estimated according to some reasonable standards of accuracy, collected in a systematic manner for a predetermined purpose placed in relation to each other."

1.5 Characteristics of Statistical Data

a) Statistics are aggregates of facts:

Single and isolated figures are not statistics. For, such figures are unrelated and cannot be compared. For instance, a single age of 40 years is not statistics but a series relating to ages of group of persons will constitute statistics.

Check your Progress

3. State the most comprehensive definitions of Statistics.

b) Statistics are affected to a marked extent by a multiplicity of causes:

Generally speaking, facts and figures are affected to a considerable extent by a number of forces operating together. For example, statistics of prices affected by conditions of supply, demand, exports, imports, currency circulation and a large number of other factors. It is very difficult to study separately the effect of each of these factors on the general price level.

c) Statistics are numerically expressed:

All statistics are numerical statements of facts, i.e., expressed in numbers. Qualitative expression like good, bad, young, old, etc., do not constitute statistics.

d) Statistics are enumerated or estimated according to reasonable standards of accuracy:

Facts and figures about any phenomenon can be derived in two ways viz., by actual counting and measurement or by estimates. Estimates cannot be as precise and accurate as actual counts or measurement. In many statistical studies perfect accuracy is not possible. The degree of accuracy depends on the object of the study and the area to be covered. For example, an estimate that five lakhs people witnessed the Republic day parade does not mean exactly five lakhs; it may be a few hundreds or thousands more or less. On the other hand, if we count the number of students, in a class and say that there are 50 students, this figure would be 100% accurate. Again, in measuring the heights of students for medical test even a fraction of a centimeter is vital. Whereas in measuring the distance between two towns, even a few metres may be ignored.

e) Statistics are collected in a systematic manner:

Before collecting statistics a suitable plan of data collection should be prepared and work must be carried out in a systematic manner. Data collected in a haphazard manner may lead to fallacious conclusions.

f) Statistics should be placed in relation to each other:

Statistics are collected mostly for the purpose of comparison. Valid comparisons can be made only if the data are homogeneous i.e., if they relate to the same phenomenon or subject. For instance, it would be meaningless to compare the height of elephant with the height of human beings.

1.6 Statistics Definition as Statistical Methods

(i) Prof.A.L.Bowley has given three different definitions of statistics. They are as follows:

(a) One of his definitions of statistics is that it is "the science of counting".

This definition is too narrow. For it takes into account only one aspect of statistics, viz., collection of data. Moreover, instead of actual counting, the estimation is made. Hence, this definition is unsatisfactory.

(b) Another definition of statistics given by Bowley is that “it is the science of averages”. This definition is also not satisfactory. For, averages are not the only devices used in statistical analysis. Besides averages, other devices such as dispersion, skewness, etc., are also used in statistical analysis.

(c) One more definition given by Bowley is as follows: “Statistics is the Science of measurement of social organism, regarded as whole in all its manifestations”. This definition is also too narrow. It limits the scope of statistics to sociology i.e., man and his activities. For, modern statistics takes into consideration not only social phenomena but also biological, astronomical and physical phenomena. Besides this, Bowley has confined the function of statistics to measurement alone while it is equally concerned with comparison, analysis, presentation and interpretation of data.

(ii) Boddington has defined statistics as “the science of estimates and probabilities”. This definition is also not satisfactory. For, estimates and probabilities are only a part of statistical methods.

(iii) Croxton, Cowden and Klein have given a very simple and comprehensive definition of statistics. To quote them “statistics may be defined as the collection, presentation, analysis and interpretation of numerical data”.

(iii) Seligman’s definition of statistics is also as simple and comprehensive as that given by Croxton, Cowden and Klein. To quote him, “Statistics is the science which deals with the methods of collecting, classifying, presenting, comparing and interpreting numerical data collected to throw some light on any sphere of enquiry.

1.7 Most Acceptable Definition

Of all the definitions of statistics given above, the definition of Croxton, Cowden and Klein and that given by Seligman are generally acceptable. For, they describe the scope and ultimate subject matter of science of statistics completely. These definitions clearly point out the four stages in a statistical investigation, viz., 1) collection of data 2) presentation of data 3) analysis of data and 4) interpretation of data.

1.8 Main Divisions of Statistics:

Any modern science is divided into a) pure science and b) applied science. For instance, the science of Economics is divided into a) Pure Economics and b) Applied Economics. In the same way, Statistics as a science has been divided into two main classes viz., a) Statistical methods and b) Applied Statistics.

Statistical Methods are concerned with the formulation of the general rules and principles applicable to the collection, classification, analysis and interpretation of data.

Applied Statistics deals with the application of those rules and principles to concrete subject matter like wages, prices, trade, population, etc. Applied Statistics may consist of biometry, (which deals with the measurement of living organism), psychometry (which deals with mental phenomena), vital statistics, administrative, social and economic statistics.

Applied statistics is divided into two classes, viz., 1) Descriptive Applied Statistics and (2) Scientific Applied Statistics.

Descriptive Applied Statistics deals with data which are known and which are related either to the present or to the past. For example, business statistics are descriptive applied statistics as they deal with the analysis, measurement and presentation of business facts, relating to the past or present. On the basis of these facts, decisions about various business problems are usually taken.

Scientific Applied Statistics deals with the formulation of scientific laws on the basis of quantitative data collected for descriptive purposes by the use of appropriate statistical methods. For instance, if we are able to make business forecasting by the use of some business statistics, we are making use of scientific applied statistics.

1.9 Functions of Statistics:

As an opera glass of telescope increases the field of physical vision, statistics increases -in the field of mental vision. That is, statistics is able to widen our knowledge. According to A.L.Bowley, "The proper function of statistics is to enlarge individual experience". Broadly speaking, statistics performs six main functions as noted below:

- (i) Condensation
- (ii) Comparison
- (iii) Correlation.
- (iv) Correction.
- (v) Measures results
- (vi) Enlarges human experience.

Condensation:

Statistics simplifies unwieldy and complex mass of data. Human mind, because of its limitations, is unable to grasp the significance of a large mass of complex data. But statistical methods make these data easily intelligible and readily understandable. The complex data may be reduced to totals, average etc. and presented either graphically or diagrammatically. These devices help us to understand the significant characteristics of the numerical data and consequently save us from a lot of mental strain. Single figures shown as averages or ratios can be grasped easily than a mass of statistical data comprising thousands of facts. In the same way, graphs and diagrams, because of their greater appeal to the eye and imagination, enable proper understanding of numerical data.

Comparison:

According to Boddington "The essence of statistics is not mere counting but comparison". Certain facts, by themselves, may be meaningless unless they are capable of being compared with similar facts at other periods of time. For instance, we collect statistics of industrial production with the primary objective of comparing them with those of the past to find out whether we have advanced or reduced, or with similar statistics collected in other countries of the world to know where we stand in the international industrial field. Some of the modes of comparison provided by statistics are: totals, ratios, graphs and diagrams.

Correlation:

Another function of statistics is to investigate the relationship between two or more phenomena. In all types of enquiries a study of the relationship between different factors is very important. For instance, the relationships between supply and demand, rainfall and agricultural yield, prices and wages, speculation and interest rate etc., require a very careful study. Statistics describes concisely the existence, direction, degree and nature of relationship or association between one phenomenon and another. The statistical measures generally used for this purpose are coefficient of association, coefficient of covariance, etc.

Correction:

In the absence of statistics our ideas are likely to be vague, hazy and indefinite. Figures help us to represent things in their true perspective. Statistics enable us to present facts in a definite and unambiguous form. For instance, if we say that wholesale prices have risen very high, our knowledge about the price remains some what vague. But when we give index numbers of wholesale prices,

our knowledge of price rise is definitely precise. Thus, statistics' puts a stamp of precision on our vague and hazy ideas. To quote Lord Kelvin, "when you can measure what you are speaking about and express it in numbers, you know something about it; when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind".

Measures Results:

Measuring the results of a particular policy is possible only with the help of statistical methods. For example, we have to find out whether a rise in the bank rate has affected the industries adversely or favourably. Here we have to compare the present situation with the past and find out whether the change has been beneficial or otherwise from the point of view of industries. Here it is impossible to arrive at any sound conclusion without the use of adequate statistical data. Statistics thus helps in measuring the effects of a particular policy and in arriving at a conclusion about it.

Enlarges Human Experience:

Statistics make it easier for man to perceive, describe and measure the effects of his own actions or the actions of others.

Scientific Applied Statistics deals with the formulation of scientific laws on the basis of quantitative data collected for descriptive purposes by the use of appropriate statistical methods. For instance, if we are able to make business forecasting by the use of some business statistics, we are making use of scientific applied statistics

1.10 Importance of Statistics

The methods of statistics are useful in an ever-widening range of human activities. That is, they are useful in any field of thought in which numerical data may be had. They are helpful to planners, industrialists, bankers, insurance companies, research, social studies, formulation of policies, etc. To quote Secrist "There are few problems relating to business, social policy or state craft for the understanding of which statistics are not required". We have explained below the importance of statistics in various spheres.

(i) Importance to State:

Statistics are very helpful to a state as they help in administration. Modern state' makes extensive use of statistical data on various problems. In fact, the

Government in most countries is the biggest collector and consumer of statistical data. They must have accurate information regarding the number of persons living within its boundaries, the total wealth of the country, extent of evils prevalent, etc. This is impossible without adequate statistics. Moreover, imports and exports, production of goods and services, labour situation in the country, price fluctuations, crimes, etc., all require statistical methods for their compilation. Again, the modern state requires numerical data for formulating economic policies, introducing social reforms and many other matters relating to day to day life. Thus, statistical methods have become quite indispensable to the modern state.

(ii) Importance in Economics:

Statistical data and statistical methods are very helpful in proper understanding of the economic problems and the formulation of economic policy.

A theoretical economist may formulate important economic laws from empirical data or the economist may verify the validity of economic laws with the help of statistical data.

Economic history can best be studied with the help of numerical facts. To find out whether a country has made economic progress or not we make use of national income statistics.

The study of economic problems is specially suited to statistical treatment. For instance, a proper appreciation of the nature and magnitude of the problem of unemployment necessitates knowledge about the following: Is unemployment increasing or decreasing? Is it widespread or confined to certain areas? Does it affect the educated and uneducated alike? etc. All these questions can be answered statistically and the data collected will enable us to form a correct estimate of the problem.

Again, statistics is very helpful in all the four branches of economics, viz; consumption, production, value and distribution. Statistics of consumption enable us to find out the way in which people belonging to different income groups in the community spend their incomes. Statistics of production help in adjusting the supply according to demand. In value, we are concerned with the pricing of goods. In price fixation, numerical data are very useful. In distribution we provide answers to questions like. How is national income calculated? How does It get distributed? These questions cannot be answered without statistics.

Check your
Progress
4. Illustrate
anyone use of
Statistics in
Economics

(iii) Statistics and Planning:

Modern age is an age of planning. Economic planning aims at maximum economic development with the available resources. Economic planning is inconceivable without adequate statistical data. If we make a study of "economic plans of country, we will find that all the plans are based on statistical data of economic activity of that country. That is, on the basis of the statistical data relating to population, agriculture, industry, trade, price, employment, etc. the planning body fixes targets to be achieved over a period of time. Again statistical methods are quite necessary to evaluate the success of the plan. To quote Tippet, "planning is the order of the day and without statistics planning is inconceivable".

(iv) Usefulness in Business:

Statistics is an aid to commerce and industry. It helps the businessmen and industrialists to decide things more precisely and objectively.

Statistical information is needed from the time the business is launched. At the time of the floatation of the concern, statistical data are required for the purpose of drawing up the financial plan of the proposed unit.

In these days of large-scale production and cut-throat competition, a manufacturer must know in advance how much is to be produced, how many workers and how much raw material will be needed to produce the estimated, quantity, and in what quality, type, size, colour or grade of the product is to be manufactured". In short, the manufacturer must have a production plan. And he cannot frame a production plan without quantitative facts. Statistics is thus a tool of production control.

Success of a business depends on accurate forecasts of sales. That is, producer or dealer should first estimate the demand for his products, analyse the possible effects of factors like changes in tastes, fashions etc. This is called business recasting. Statistical methods are very helpful in business forecasts. For instance, a forecast of sales can be made by studying time series of present and prospective conditions of the concern and related business.

Statistics is also useful in conducting market research by businessmen. A skillful analysis of data on population, purchasing power, customs of people, competition, transportation costs, etc., should be made before launching a new product. Again, in by\ping up and maintaining an extensive market, it is important to keep accurate records of its present and potential geographical distribution.

Again cost accounting is entirely statistical in outlook and it is with the

Check your Progress

4. Illustrate any use of Statistics in Economics

help of this technique that producers are in a position to decide about the prices of various commodities.

Space for hints

Statistical quality control is a standard procedure for improving the quality of manufactured goods. The number of defective articles produced by a machine is recorded continuously on a chart, and the statistician, looking at the chart, can tell when the production is becoming worse, and thus he maintains the quality of the manufactured articles at a uniformly high level.

Statistics is also very helpful in business management. With the growing size of business enterprises, centralised management of all the activities of an industrial concern from production to sales has become impossible. Separate business departments are to be created on the basis of function they serve such as planning, material control, purchasing, sales promotion, accounting, personnel etc. For the proper execution of these functions, the management has to depend upon statistical data and accounting records.

In fact, at present, many business concerns have got a separate statistical department.

(v) Utility to Insurance Companies:

Statistics is extensively used in the field of insurance. The entire working of life insurance schemes rests on the compilation of life tables and computation of expectation of life from time to time. Similarly, unemployment insurance and sickness insurance depend on statistical data. The theory of Probability finds its full use in the field of insurance. In short, insurance as an institution could never have existed in the absence of statistical data.

(vi) Usefulness to Public Utility Concerns:

Public utility concerns like railway, electricity supply companies, water works etc., also use statistics extensively. Statistical analysis of railway working is every useful in the expansion programmes. Moreover, railway companies are facilitated in running special and additional trains on some routes on the basis of their estimated traffic. A water work is in need of a list of users of water for various purposes to fix differential rates.

(vii) Usefulness to Bankers:

Statistical methods are also useful to bankers and stock exchange brokers. A banker who intends to build up a pyramid of credit should have adequate knowledge of seasonal variations in demand for bank credit. A stock exchange broker or an investor in securities needs a knowledge of seasonal variations in

securities, needs a knowledge of interest rate, the fluctuation of investment market and other related data to strike a timely bargain.

(viii) Statistics And Research:

Statistics is an indispensable tool of research. Now-a-days statistical methods are being used in research in almost every sphere. For example, experiments about crop yields and different types of soils or the growth of animals under different diets and environments are very often designed and analysed according to statistical methods. Even in the fields of medicine and public health, statistical methods are used for testing the efficiency of new medicines and new methods of treatment.

(ix) Importance in Social Studies:

Statistics is indispensable in social studies. A social worker has to rely upon statistics for carrying on his activities in right direction. The magnitude of beggar problem, the extent of child-marriage, magnitude of illiteracy and unemployment, etc., all require a careful statistical treatment for tackling them properly.

(x) Helpful for the Formulation of policies:

For framing suitable policies statistics is very helpful. For instance, it may be necessary to decide how much wheat India should import next year. The decision would depend upon the expected domestic production and the likely demand for what in next year.

(xi) Usefulness in Other Fields:

Statistical methods are applicable wherever quantitative studies are to be made. In Political Science they are used to ascertain public opinion through the random sampling technique. In the field of education, statistical methods are used to conduct the aptitude test rating known as intelligence quotient (I.Q) In Astronomy, the method of least squares, an important statistical technique is used to trace the paths of stars, comets, planets and other heavenly bodies. In biological studies, the method of correlation analysis is widely used for testing hereditary characteristics. Vital statistics are of very great help to the science of medicine. The principle of probability finds wide application in engineering. In agriculture, to assess the yield of the different crops and quality of produce, statistical methods are used.

Thus, statistical methods and studies are of universal application. Nowadays there is hardly any field where statistical methods are not used. Hence, it has

been pointed out, “statistics affects everybody and touches life at many points”. Bowley’s statement highlighting the importance of statistics is worth quoting. It is as follows: “A knowledge of statistics is like a knowledge of foreign languages or of algebra; it may prove of use at any time under any circumstances”. Recently, use of statistical methods is fast increasing. Indeed, the prophecy of H.G. Wells, viz. “Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write” - has come true.

1.11 Role of Statistical Methods in Other Fields With Special Reference to Economics:

The methods of statistics are useful in an ever widening range of human activities. To quote Croxton, Cowden and Klein, “The methods of statistics are useful in an ever widening range of human activities in any field of thought in which numerical data may be had. “In modern times, there is no science where statistics does not find application. Statistical methods find application in political science, biology, medical science, meteorology, agricultural science, physical science, astronomy, engineering, mathematics, econometrics and economics. In view of its applicability in almost all the disciplines it has been stated as follows:

Sciences without statistics bear no fruit;

Statistics without sciences has no foot.

We have explained below the relationship between statistics and other disciplines.

(i) Statistics and Political Science:

In Political Science, statistical methods are used to ascertain public opinion on important social, economic and political issues. These kinds of opinion are important for running the administrative machinery of the government, efficiently. In America, there is Gallop Poll which has been ascertaining public opinion (by random samples) on matters of topical interest in politics, social field etc., from time to time.

(ii) Statistics and Biology:

In all branches of Biology, methods of statistics are used. According to Karl Pearson, the whole doctrine of heredity rests on statistical basis. Sir Francis Galton conducted a statistical enquiry into the principles relating to the adoption of mental and physical characteristics from one generation to another. The method of correlation analysis is applied to study the correlation between age of husband and that of wife and between height of father and that of son. Again, experiments

about the growth of animals under different diets and environments are frequently designed and analysed according to statistical principles.

(iii) Statistics and Medical Science:

Statistics plays a useful role in the medical science. Statistical methods are applied to study the potency of drugs to cure diseases and to find out the safety limits for usage. Hospitals keep medical records of patients. Death rates and sickness rates of people are recorded to study epidemiology.

(iv) Statistics and Meteorology:

The science of statistics help meteorology in a large number of ways. In meteorology records are made of temperature, humidity of air and barometrical pressures. For purposes of comparison and forecasting it becomes necessary to average these figures and to study their trend and fluctuations. All these cannot be done without the use of statistical methods.

(v) Statistics and Agricultural Science:

The collection of Agricultural statistics is a major activity of the Revenue and Agriculture Departments of the Government. Crop forecasts, which are very useful for the trade and as a stabilizer of prices, are done with the help of statistical methods. Correct statistics are also necessary to bring in more areas under plough, more areas to be irrigated and for the formulation of cash crop, food crop policy. Moreover, layouts, designs of experiments, effect of application of fertilisers on yield of crops and quality of the produce, area affected by pests, etc., are worked out with the help of statistical techniques.

(vi) Statistics and Physical Sciences:

The exact sciences like Physics and Chemistry have benefitted to a great extent by statistics. In Physics, observations taken to find out speed of light give nearly equal answers. To find out the limits of the correct answer, statistical methods are used. Again, the behaviour of atoms, electrons, etc., in general are not exactly alike. They have different speeds and direction. Hence, a study of their behaviour is made through the formulation of statistical laws.”

(vii) Statistics and Astronomy:

Statistics were first collected by astronomers for the study of the movement of stars and planets. They applied the statistical methods for the furtherance of their studies. In fact, the method of least squares was first developed by an astronomer. This method is an important statistical technique and is still applied to trace the paths of stars, comets, planets and other heavenly bodies.

(viii) Statistics and Engineering:

Statistical methods are also used by the engineering science. Many problems in Engineering are solved by using the principle of probability. Control of quality and standardisation of manufactured products time study in machine operation- all require knowledge of statistical methods.

(ix) Statistics and Mathematics:

Statistical methods constitute a branch of mathematics in as much as it is with the help of mathematical formula that the data are analysed, compared and interpreted. Averages, measurement by deviations, co-efficients, construction of index numbers, etc., have their foundation in mathematics. Again, it is the mathematical theory of probability which furnishes the basis for the law of statistical regularity on which it raised the entire supersructure of the modem theory of statistics. In fact, statistics is a part of Applied Mathematics. To quote W.I.King, Statistics may properly be "Considered as a branch of mathematics in as much as it attempts to formulate definite rules, procedures applicable in handling groups of data of many different varieties".

Although statistics is a branch of mathematics, it is not as exact science - as mathematics. Bowely has drawn a line between mathematics and statistics as follows; "whereas arithmetic attains exactness statistics deals with estimates, very accurate and often sufficiently so for their purpose, but never mathematically exact".

(x) Statistics and Econometrics:

Econometrics is a science of recent origin. It is a science that has been evolved through a fusion of Economics, Mathematics and Statistics. The science of econometrics provides solution to economic problems through the use of economics, statistics, and mathematics. The interrelationship among these sciences of Economics, Mathematics and statistics resulting in the evolution of the science of Econometrics has been very helpful to the development and progress of all these three sciences.

(xi) Statistics and Economics:

Statistical data and statistical methods are very helpful in the proper understanding of the economic problems and that of formulation of economic policy.

A theoretical economist may formulate important economic laws from empirical data. Or the economist may verify the validity of economic laws with

the help of statistical data. To quote Karmel and Polask "Statistical methods make possible to the development of the empirical side of economics. Their use is necessary to give real content to theoretical formulation".

Economic history can best be studied with the help of numerical facts. To find out whether a country has made economic progress or not we make use of national income statistics.

The study of economic problems is specially suited to statistical treatment. For instance, a proper appreciation of the nature and magnitude of the problem of unemployment would necessitate knowledge about the following: Is Employment increasing or decreasing? Is it widespread or confined to certain areas? Does it affect the educated and uneducated alike? etc. All these questions can be answered statistically and the data collected will enable us to form correct estimated of the problems.

Statistics is very helpful in all the four branches of economics, viz., consumption, production, value and distribution. Statistics of consumption enable us to find out the way in which people belonging to different income groups in the community spend their incomes. Statistics of production help in adjusting the supply according to demand. In value, we are concerned with the pricing of goods. In price fixation, numerical data are very useful. In distribution, we provide answers to questions like – How is national income calculated? How does it get distributed? These questions cannot be answered without statistics.

Recently, a new science called Econometrics, which is a fusion of Economics, Mathematics and statistics has come into existence. It has become very popular. It has been very helpful for the solution of economic problems.

To test the hypotheses of pure economic theory and the conclusion drawn from these hypotheses, empirical investigations involving statistical data are needed. Hence, Marshall has made the following observation, "statistics are the straw out of which I, like every other economist, have to make the bricks".

1.12 Limitations of Statistics

At present, statistical methods are being used in almost all the fields of scientific investigation. Hence, it has been pointed out that science without statistics bear no fruit. Despite the universality of its application, the science of statistics has its own limitations. In the words of Rhodes "Statistics, while an extremely useful tool to the investigator in almost in any line of scientific inquiry has its limitations and shortcomings". The important limitations of statistics are considered below:

(i) Statistics cannot take cognisance of individual items:

Statistics deals with aggregates of facts. Hence, the study of an individual fact lies outside the scope of statistics. In the words of W.I.King, "Statistics from the very nature of the subject cannot and never will be able to take into account the individual cases". Hence, where knowledge about individual cases is essential, statistics proves inadequate. For instance, the per capita income of people in the country might look quite satisfactory. From this we should not conclude that there are no people in the country getting very low incomes. For, there may be people getting fabulous as well as very low incomes. Hence, from the per capita income figure, we should not jump to the conclusion that each and every person in that country is earning fairly good income. Thus, statistics fails to throw light on the real position.

(ii) Statistical results are true only on an average:

That is, the laws of statistics are ~ not universally true like the laws of physics. In the words of L.R.Connor, "Statistical law is only held to be true on an average or in the long run". For instance, average rainfall in a city may be 40 "per year. This does not mean that in any year rainfall will not be more or less than 40". But in the long run the average rainfall will be 40". Thus, statistical laws are true on the average or in the long run.

(iii) Statistics can study only the quantitative aspect of any problem:

All statistics are numerical statements of facts. Hence, qualitative characteristics like honesty, efficiency, intelligence, blindness, deafness, etc., cannot be studied directly by the use of statistical methods. However, it may be possible to analyse such problems statistically by expressing them numerically. For example, we can study the intelligence of boys on the basis of the marks secured by them in an examination.

(iv) Statistics is only one of the methods of studying a problem:

On several occasions statistical methods cannot provide complete solution for the problem under investigation. For instance, when we study the economic set up of a country statistically, it is to be supplemented by other evidences like the country's cultural and religious background.

(v) Statistical data must be uniform and homogeneous:

With out homogeneity of statistical data comparisons would be difficult. Let us consider an example. Suppose we have to compare the wages in two establishments. In one establishment, the average wage is composed of adult wages only. But in the other establishment, the average is composed of wages of

adults and children. In that case, it is not possible to compare the wage structure of these establishments.

(vi) Statistical results are generally estimates rather than exact statements:

For, when data are collected through interview or observation, there is large scope for error due to bias on the part of the interviewer or the interviewee. Hence, L.R. Connor has made the following remark: "Statistical data must always be treated as approximations or estimates and not as precise measurements".

(vii) False conclusions might be arrived at if statistics are quoted without the context:

An example will enable us to understand this point. Suppose we are given the information that the average marks secured by two students in three tests are 50 per cent. On the basis of this statistical data we may conclude that both have made lential progress. But when we are told that one student has been petting 40, 50 and 60 percent and the other 80, 40 and 30 percent in the three successive tests, we will have to reverse our judgement. That is, while the progress of one of them is positive, that of the other is negative. Thus, as L.R. Connor has stated, "Statistics provide a basis for judgement but not whole judgement".

(viii) Statistics are liable to be misused:

Only those who possess an expert knowledge of statistical methods can scientifically handle statistical data. Like medicines in the hands of quacks, statistics are capable of being easily misused by the inexperts. Statistical data may be wrongly used without knowing the purpose for which and the circumstances under which they were collected. Ignorance of the meanings and definitions of the terms used and of the degree of accuracy of the data may lead to wrong conclusions. They may also be used by unscrupulous investigators to mislead the public. Hence, as Bowley has pointed out, "Statistical methods are most dangerous tools in the hands of inexperts".

1.13 Distrust of Statistics:

In spite of its high esteem and application in every sphere of life there is some amount of misgivings in the minds of some persons with regard to the usefulness and reliability of statistics. They have given expression to their feeling in a number of ways. Following are the often quoted examples:

1. "There are three kinds of lies—lies, damned lies and statistics (Mark Twain)
2. Statistics are lies of the first order.
3. Statistics can prove anything.

4. Statistics are like clay of which one can make a god or devil as one likes.

Space for hints

5. An ounce of truth will produce tons of statistics.

6. If figures say so it cannot be otherwise.

7. Statistics are white lies.

Such misgivings are due to the following factors:

1. Figures always carry conviction and therefore people are easily led to believe them.

2. Figures are capable of being easily manipulated to serve one's own interests.

3. Even if correct figures are used, they may be presented by selfish and unscrupulous persons in such a manner that the public is misled. For example, a labour leader in India may tell the workers that the workers in America are paid high wages. And he may give correct figures of average wages paid in American industries. But he conceals the cost of living and other factors for which a higher wage is necessary in America. The data are incomplete and are deliberately manipulated.

Prof.A.L.Bowley has pointed out the following as causes of distrust of statistics.

1. Figures may be quoted without their context.

2. Figure may be applied to a group of phenomena quite different from that to they are related.

3. The estimate relating to only part of the group may be taken as complete.

4. Only those facts favourable to the argument may be given.

5. Sometimes it is argued hastily froth effect to cause.

Besides the factors mentioned above there are other factors responsible for the mistrust of statistics. They are noted below:

1. It is not easy to distinguish between reliable and unreliable statistics. In the words of W.I.King, "one of the shortcomings of statistics is that they do not always bear on their face the label of their quality. The crudest table, founded on the most unreliable basis appear's to the casual oberver equally valuable with a table compiled after some months of labour by a corps of skillful statisticians.

2. Inaccuracy of data and mistakes in calculation.

3. Wrong conclusions based on false assumptions.

4. Deliberate twisting of facts.

It is to be noted that distrust of statistics is mainly due to lack of knowledge regarding the nature, limitations and utility of statistics and not due to any basic demerit of the science of statistics. Statistics neither proves any thing nor disproves anything. For, it is only a tool, i.e., a method of approach. Tools, if properly used, do wonders and if misused, prove disastrous. For instance, a very sharp knife is very helpful to a gardener but it is a dangerous toy for the kids. Again, medicine in the hands of expert doctors, will be most useful. But, if administered by a quack, it may prove fatal to the patient. That is true of statistical tools. It is to be noted that "figures do not lie, but liars can make figures lie". If properly used, statistical tools help in making wise decisions and if misused proved disastrous. As A.L.Bowley remarks, "statistics only furnish a tool necessary, though imperfect which is dangerous in the hands of those who do not know its use and deficiencies. Statistical methods are more dangerous in the hands of inexperts". To conclude in the words of W.I.King; "Statistics is the most useful servant but only of great value to those who understand its proper use".

2. PLANNING AND CONDUCTING A STATISTICAL ENQUIRY

2.1 Meaning of Statistical Enquiry:

The term 'enquiry' means a search for knowledge. Therefore a statistical enquiry implies search for knowledge through statistical device about some problem. The end result of such a research is figures. These figures enable the analyst in arriving at certain conclusions. Suppose, it is desired to know about the consumption pattern of a particular class of people. In that case, arrangements are to be made for conducting a statistical enquiry relating to the consumption pattern of that class of people. This will necessitate collection of data by the investigator relating to the percentage of income spent by these people on food, clothing, shelter, education, fuel and lighting, etc. These data will enable the analyst in arriving at conclusions regarding the consumption pattern of these people.

2.2 Types of Statistical Enquiry:

Statistical enquiries are of several types. The most important types of statistical enquiries are listed below. An enquiry may be

(i) General purpose or special purpose;

(ii) Direct or indirect;

(iii) Confidential or open;

(iv) Original or repetitive;

(v) Regular or ad hoc;

(vi) Census or sample;

(vii) Mail-card enquiry or enquiry through enumerator;

(viii) Official or semi-official or non-official;

General purpose enquiry :

In this type of enquiry, we collect data which are useful for several purposes. An example of this type of enquiry is the population census taken in every 10 years in our country. Such an enquiry provides information not only about the total population but also about its division into males and females, literates and illiterates, employed and unemployed, age distribution and income distribution. In special purpose enquiry, we collect data which are useful in analysing a particular problem.

Direct Enquires:

Direct enquires are those in which the data are capable of quantitative expression. For instance, height of students is capable of direct quantitative measurement. Hence, a statistical enquiry relating to the height of students is an instance of direct enquiry.

Indirect Enquires:

In an indirect enquiry, direct quantitative measurement is not possible. But some related information expressible in numerical quantities and bearing indirectly on the subject matter is collected. For instance, to conduct a statistical enquiry relating to the intelligence of students the marks obtained by the students in certain examination may be collected and used.

Confidential Enquires:

Here the results of the enquiries are kept secret and are not made known to the public. Generally enquiries conducted by Trade Associations and Chambers of Commerce are of this type.

Open Enquires:

In the case of open enquiries, the results of the enquires are made known to the general public.

Original Enquires:

An original enquiry is one which is carried out for the first time. Hence, in the case of original enquiry a plan suitable for the purpose will have to be made and work will have to be undertaken from the beginning.

Repetitive enquiry:

It is conducted in continuation or repetition of previous enquiries. Hence, in the case of a repetitive enquiry, the old plan with such modifications as experience or necessity demands, is followed.

Regular Enquires:

In the case of regular enquiries, data are collected at regular intervals over a period of time. For instance, for constructing an index numbers series, prices of different commodities may be collected every week.

Ad hoc enquiries:

In the case, data are collected as and when necessary without any regularity. That is, an ad hoc enquiry is conducted once in a while. For instance, an enquiry into the extent of begging in a country is conducted only occasionally.

Census enquiries:

In a census enquiry, every individual or object of enquiry is surveyed e.g. the census of population.

Sample enquiries:

In sample enquiry, some selected representative individuals or objects pertaining to the field of enquiry are studied. The National Sample Survey is an example of such enquiry in India.

Mail-card enquiries:

In the case of mail card enquiries, a printed form known as questionnaire is sent by post to each informant. The informant is asked to fill it out and return the filled up form to the investigator by post.

Enquiries through enumerator:

In this type of enquiry, trained enumerators are appointed. They approach the informants, interview them and record the informations supplied by them in forms known as schedules.

Official enquiries:

Enquiries conducted by on behalf of Central or State Governments are called official enquiries.

Official enquiries:

Enquiries conducted by such bodies enjoying government patronage (for instance, Universities) are called semi-official enquiries.

Non-Official enquiries:

Enquiries conducted by private bodies or individuals are called non-official enquiries.

2.3 Factors to be taken note of while planning the enquiry:

Before the actual collection of data, a well thought out plan has to be prepared so that money, time and labour may not be wasted over wrong methods. The matters that require careful consideration at the planning stage are noted below:

(i) Purpose of Enquiry:

The purpose for which the statistical enquiry is to be conducted must be clearly and precisely laid down. This will indicate the type of information which is needed and the use to which the information obtained will be put. For example, if the object of an enquiry is to study the nature of price change over a period of time. It would be necessary to collect data on commodity prices and it must be decided whether it would be helpful to study wholesale or retail prices, and the possible uses to which such information could be put. This will help the investigator to collect the proper information reducing wastage of time and money.

(ii) Scope of enquiry:

The Scope of enquiry must also be spelt out clearly and unambiguously. The scope of the enquiry usually fixes the limits of the enquiry. (a) The nature of information to be collected (b) the geographical area to be covered (c) type of enquiry and (d) the time limit for completion of the enquiry are to be clearly indicated.

(iii) Determination of Units of Data Collection:

The statistical approach to any problem is based on measurement or counting. Individuals or objects that form the subject-matter of enquiry are enumerated with reference to some attributes. For instance, we might investigate 'accident', 'income', 'employment', 'wage' etc. These are statistical units.

Physical units of measurement like ton, pound, yard, year, etc., are well determined and defined. These units do not need any explanation or definition. However, in many statistical studies such customary and legal units are not available. In such, cases, the statistician has to arbitrarily decide about a unit and has to give 'its proper definition. Suppose an enquiry is to be conducted about the wages of workmen in some industry, Wage is a common term which may mean money wage or real wage, piece wage or time wage, wage of skilled worker or wage of unskilled worker, etc. One who conducts the enquiry must state clearly the sense in which he proposes to use the term 'wage'.

The statistical unit for any enquiry must be determined with caution and care. Otherwise, things which should be omitted might be included, while things which ought to be included might be omitted.

Essential requisites of a statistical unit:

- (a) It must be simple, clear and explicit.
- (b) It must be definite, specific and ascertainable.
- (c) It must ensure homogeneity and uniformity. That is, the unit should not mean different things at different occasions.
- (d) It should be a stable and a standard unit.
- (e) I It must be suitable for the object of the enquiry undertaken.

Classification of statistical units :

There are two ways of classification of statistical units. They are 1) spontaneous units, 2) artificial units.

The spontaneous units are those used in counting process. They are of two kinds, viz., (a) natural units and (b) produced units. Examples of natural units are a person, a tree, a cow, etc. By produced units we mean natural materials, changed for human advantage and uses; examples of produced units are a house, a car, a table etc.

The artificial units are those used in measurement process. They are of two kinds viz., (a) mensurational units and (b) produced units. Examples of mensurational units are the ton, the metre, the year; etc. Units used, for measuring values are called pecuniary value units. Examples of pecuniary value units are rupee, dollar, sterling, etc. The pecuniary value units are indispensable for measuring the value of production, trade, etc.

Classification of statistical units according to their functions:

Space for hints

- (a) Units of collection
- (b) Units of analysis and interpretation.

Units of collection:

They are known as units of enumeration or estimation of Units of collection are those in terms of which measurements are made or data collected. These units are of two types, viz (a) simple and (b) composite. A simple unit expresses a single condition without any qualification. Example of simple units are a worker, a mile, a building etc. A composite unit is one which is formed by adding a qualifying word to a simple unit. For instance, a worker is a simple unit but industrial worker is a composite unit; a mile is a simple unit but ton~ mile becomes a composite unit. A simple unit is a unit which is in common use and hence is not difficult to define it. But the definition of a composite unit becomes difficult by the addition of some qualifying condition to a simple unit.

Units of analysis and interpretation:

They are those which make comparison possible and easy. These units include (a) rates (b) ratios (c) coefficients.

(iv) Sources of Data Collection:

There are two sources of information available to the statistician. They are (a) primary source and (b) secondary source. If the investigator collects first hand data for the purpose at hand, such data are known as primary data. Instead, if he gets the data from published or unpublished sources, such data will constitute secondary data for him.

(v) Technique of Data Collection:

There are two important techniques of that collection namely, (a) census technique and (b) sample technique. In a census enquiry, every individual or object connected with the subject or enquiry is surveyed e.g., the census of population. In a sample enquiry, some selected representative individuals or objects pertaining to the field of enquiry are studied. The National Sample Survey is an example of such an enquiry in India. The census method is costlier and more time consuming as compared to the sample method. The investigation must decide which technique he will use. The choice would depend upon a number of factors such as (a) the availability of resources (b) time factor (c) the degree of accuracy desired and (d) the nature and scope of the problem.

Check your Progress

5. State the various types of statistical units.

6. What are the sources of statistical data?

(vi) Choice of Frame:

Having set the limits of the population or whole universe of enquiry, it is essential to identify the units which constitute the population. A list of the units which constitute the available information relating to the subject of enquiry is called the frame of the enquiry. Suppose we want to find list the capital invested and number of workers working in small-scale industries in Madurai. For this purpose, we must have a complete list of names and addresses of all the small scale firms in Madurai. The list of names and addresses will be the 'frame' for this enquiry. The whole structure of enquiry is to a considerable extent determined by the frame.

(vii) Degree of Accuracy Desired:

An important step in planning a statistical investigation is the determination of the standard of accuracy that is to be observed in the collection of statistical material. Absolute accuracy or mathematical exactness in statistical investigation is neither possible nor necessary. But statistical data must possess a reasonable degree of accuracy and every effort should be made to attain it. Degree of accuracy desired primarily depends upon the object of enquiry. For instance, in measuring the heights of students for a medical test, even a fraction of an inch is vital, whereas, in measuring the distance between two towns, even a few yards may be ignored.

(viii) Cost of the plan:

While executing the plan of enquiry, money has to be spent on the different stages of enquiry. Hence, an estimate of the cost of the survey under such headings as preliminary work, field investigation, tabulation, analysis, etc., is to be made. A preparation of cost estimates will ensure avoidance of wastage of resources.

2.4 Executing the plan of enquiry:

After the preparation of the plan of data collection, the next step is the execution of the survey. The various steps that should follow the preparation of the plan of data collection are as follows.

1. Setting up of an administrative organisation.
- 2 Design of forms and questionnaire.
3. Selection and training of the field investigators.
4. Supervision of field work.
- 5 Arrangement of follow-up in case of non-response.

6. Presentation of information.

7 Analysis of collection of data

8 Preparation of report

Space for hints

3. COLLECTION OF DATA OR SOURCES OF DATA

3.1 Sources of statistical data :

Collection of data is the first step in any statistical investigation. Statistical data are of two types.

(i) Primary data

Primary data are those which are collected for the time and are original character. Thus, the primary data are first hand data.

(ii) Secondary data

Data which have already been collected, tabulated and presented in some form by some one use for some purpose are called secondary data. As its name itself conveys, it is second hand data.

It is said that primary data are in the shape of raw materials to which statistical methods are applied for the purpose of analysis and interpretation whereas secondary data are in the shape of finished products since they have been treated statistically in some form or the other.

The distinction between primary data and secondary data is one of degree only. . Data which may be primary for one agency may be secondary for the other and vice versa. For instance, the data collected during census (of populations) operations are primary to the census department of the Government of India, but to a person which makes use of these data for further research, they will be termed "secondary".

On the basis of the classification of data made above we have two methods of data collection, viz.

1. Primary Method.

2. Secondary Method.

3.2 Primary data versus secondary data:

Primary data are (a) truthful and (b) purposive. However, primary data have two advantages compared to secondary data. They are : (1) It takes time to collect data by the primary method (2) Its collection involves much expenditure of money.

Check your Progress

7. What is primary data?

8. What is secondary data?

The investigator has to decide at the outset whether he proposes to use primary data or secondary data in his investigation. The choice between the two depends on the following considerations:

1. Nature, object and scope of the enquiry ;
2. Availability of finance;
- 3 . Availability of time; and
4. Degree of accuracy desired.

3.3 Methods of Collecting Primary Data:

(a) Direct Personal Investigation Method:

In direct personal investigation, the investigator has to collect the information personally from the sources concerned. He has to be on the spot for conducting the enquiry and he has to meet people from whom data have to be collected. For instance, if a person wants to collect data about the wages of workers of 'the Madura Mills, he has to go to the Mill, contact the workers and obtain the required information.

Advantages of the method:

1. The data collected through this method are likely to be more reliable and accurate because the investigator himself collects the information.
2. In this method, the investigator can always keep in mind the object, scope and nature of the enquiry.
3. Personal visit may ensure higher degree of co-operation from the informants.
4. Under this method the language of communication best suited to the status and educational level of the person can be adopted.
5. This method is very useful when the scope of the enquiry is limited and the investigator has sufficient time to deal with individual items.

Disadvantages of the method:

1. It may be very costly where the number of persons to be interviewed is large and they are spread over a wide area.
2. The success of this method depends largely on the personal qualities of the interviewer, his tact, diplomacy, courage curiosity, etc., which only few possess.

3. In this method, the bias or prejudice of the investigator may do a lot of damage as he is in sole charge of data collection.

4. The time required for collecting information is likely to be more under this method than in other methods.

This method of data collection is suitable where the field of enquiry is limited and within the access of the investigator.

(b) Indirect Oral Investigation Method :

In this method, enquiry is made through enumerators specially appointed for the purpose. The enumerators, instead of directly approaching the informant, interview several third persons who are directly in touch with the information sought. The third persons interviewed are known as witnesses. For example, in an enquiry regarding addiction to alcoholic drinks, people may be reluctant to supply information about their own drinking habits. In that case, the desired information can be got from the dealers of liquor or other people who may be knowing them, for example, the neighbours, friends, etc. This method is generally adopted by enquiry committees or commissions appointed by the Government.

This method is very common in economic and social surveys. The success of this method depends upon the following factors:

1. The type of person whose evidence are being recorded.
2. The personal knowledge of the witnesses about the things asked.
3. The ability of the interviewers to draw out the information from witness by means of appropriate questions and cross examination.
4. The honesty of the interviewers entrusted with the task of the collection of information.

According to L.R.Connor, "This method is useful when the information desired is complex or there is reluctance or indifference on the part of informants".

(c) Information through local correspondents:

Under this method, the investigator appoints local agents or correspondents in different places to collect information. They are advised to use their own judgement as to the best way of obtaining it. For example, in the construction of wholesale price index numbers, regular information is obtained from correspondents appointed in different areas. This method is also in wide use in the case of crop estimates.

Obviously, the data collected through this method cannot be very reliable. As such this method is suitable only in those enquiries where the purpose of study can be served with rough estimates only and where a high degree of accuracy is not essential. In the words of L.R. Connor, "This method is useful when figures are required cheaply and expeditiously and accuracy is not of prime important?".

(d) Mailed Questionnaire Method:

Under this method, a list of questions pertaining to the enquiry known as a questionnaire is prepared. The questionnaire contains questions and provides space for answers. The questionnaires are sent by post to the informants. A request is made to the informants through a covering letter to fill up the questionnaire and send it back within a specified time.

The success of this method depends upon the following factors:

- (i) The ability with which the questionnaire is prepared.
- (ii) The knowledge of informants about the facts wanted.
- (iii) The favourable response from the informants.

Merits of the method:

1. A large field of enquiry can be covered very easily.
2. It is relatively cheap and expeditious provided the informants respond in time.

However, this method is subject to certain limitations as noted below:

1. This method cannot be used to elicit information from illiterate people.
2. In the absence of personal contacts, there is nobody to explain and clarify the meanings of doubtful questions.
3. The information supplied by the informants may not be correct and it may be difficult to verify the accuracy.
4. There is possibility of non-response on a large scale. For instance, Dr. V.K.R.V. Reid; in his enquiry about Wages in estimating the national income of India in 1931-32, sent 8,143 letters to various informants. But he could receive only 130 replies containing forms duly filled in.

This method is best suited where the field of investigation is very vast and the information are spread over a wide geographical area.

(e) Schedules sent through Enumerators:

Space for hints

Under this method, a team of enumerators is selected and trained. They are provided with schedules. The enumerators go to the informants along with the schedule. They get replies to the questions contained in the schedule and fill them in their own handwriting in the schedule.

Enumerators are to be selected with due care keeping in mind the following points:

1. They should be honest, unbiased, intelligent and hard working.
2. They should evince keen interest in investigation.
3. They should be given proper training.
4. They should be tactful and be of a pleasant disposition.
5. They should be well-acquainted with the local customs.

Important advantages of the method :

1. In those cases where the informants are illiterate, this method can be adopted.
2. The problem arising from non response is minimised under this method.
3. Questions which are ambiguous can be explained clearly to the informants. Moreover, answers can be checked on the spot itself by cross examining the informants. Hence, the data collected will be more reliable and accurate.

Disadvantages of the method :

1. Of all the methods of collecting primary data, this method is the costliest as enumerators are generally paid persons.
2. The time taken in collecting data by this method is longer as the enumerators have to meet the respondents at the latter's convenience.
3. The success of this method depends largely upon the calibre of and training imparted to the enumerators.

This Method is best suited in a large scale enquiry like conducting population census.

3.4 Choice of the best method to collect primary data:

Of the several methods of data collection discussed above it is difficult to say which one is the best method. Each method has its own merits and demerits. A.L Bowley is of the opinion that in collection of statistical data, common sense is the chief requisite and experience, the chief teacher". However, while choosing a particular method of data collection, it should be done bearing in mind the following points.

1. Nature of enquiry
2. Aim of the enquiry.
3. Scope of the enquiry.
4. Availability of funds and time and
5. Degree of accuracy desired.

3.5 Design of Questionnaires and Schedules:

Questionnaires and schedules are used in the collection of primary data. By a questionnaire or schedule we mean a list containing questions pertaining to the enquiry. The questionnaire or schedule contains questions and provide space for answers.

The Questionnaire differs from the schedule in one important respect. The Questionnaire is sent by post or delivered to the informant in some other way. And the questionnaire is filled up by the informant unaided by the presence of the investigator or enumerator, whereas, a schedule is filled up by the enumerator who can interpret the questions wherever necessary.

The questions that find a place in a questionnaire or schedule have been classified into four important types as noted below:

1. Specific Information Questions:

The questions require specific answers. Examples of this type of questions are as follows: What is your age? How many children do you have?

2. Open Type Questions:

The Questions do not require Specific answers. The informants are free to give any reply in their own words to these questions. Examples of the type of questions are as follows: What is your opinion on India's Eighth Five year Plan? What is your opinion on Indian family planning Programme?

3. Alternative Type Questions:

These questions are to be answered in a word like 'yes' or 'no', 'for' or 'against' 'true' or 'false'. In case the informant is not able to take sides, he can say 'Don't know' which will be recorded as 'DK'.

4. Multiple Choice Questions:

In this type of questions answers in the form of all possible alternatives are given; and the respondent is asked to put a mark against the answer considered appropriate by him. For instance, in a question regarding marital status, the alternatives are,

Single _____

Married _____

Widowed _____

Divorced _____

Separated _____

The success of the questionnaire method of collecting information depends largely on the proper drafting of the questionnaire. Construction of a suitable questionnaire is a highly specialised job requiring great deal of skill and experience. There is no hard and fast rule to be observed in drafting the questionnaire. However, the following general principles may be helpful in framing a questionnaire.

1. The person conducting the survey must introduce himself and state the objective of the survey. A short letter stating the purpose of the survey may be enclosed. This letter may contain an assurance to the effect that the answer furnished will be kept in strict confidence.

2. The questions should be clear, unambiguous and precise. The questions should be capable of being answered only in a limited number of ways. Croxton, Cowden and Klein have given 'an example of ambiguous question. In a questionnaire sent by an organisation to hundreds of parents, the following ambiguous question was found. 'Is your child's outlook on life broader or narrower than yours was at the same age?'

The investigator expected the replies to this question to read 'Broader' or 'Narrower'. But actually replies received were (1) Yes (2) No (3) I doubt it (4) I hope so, which has no meaning.

Space for hints

Ambiguous questions or questions that invite ambiguous answers produce useless data and involve waste of time and money. Hence, as Croxton, Cowden and Klein have remarked, "The investigator should not be satisfied merely with wording his questions so that they can be understood; he should draft them so carefully that they cannot be misunderstood"

3. When alternate type questions are asked, the listed categories should be exhaustive and mutually exclusive. For instance, in an enquiry concerning marital status the question in the form of 'Married or Single' is not exhaustive. The proper way of asking this question will be checked whether,

Single _____

Married _____

Widowed _____

Divorced _____

Separated _____

4. Leading question (i.e. question which suggest the answers) must be avoided. For example, in an enquiry conducted during a period of depression question should not be of the following form:

"By how much has your income fallen?" Instead, it should be of the following form:

"Has your income fallen?" State YES or No

"If yes, by how much?"

5. Certain types of questions should be avoided. Questions which are liable to offend should be avoided. Again, questions of a personal and pecuniary nature should not be asked. For example, questions about source of income may not be willingly answered in writing. Where such information is essential, it should be obtained by personal investigation.

6. Question should not require calculation to be made. For example, informants should not be asked yearly income, for, in most cases they are paid monthly. Similarly, questions necessitating calculation of ratios and percentages should not be asked as 'it may take much time to calculate them and the informant may not send back the questionnaire at all.

7. As far as possible the question should be of such a nature that they can be answered briefly in YES or No; or in terms of numbers, place, date etc. If there are large number of questions requiring lengthy answers, the informants may not send replies to them simply for lack of time.

8. Questions should be capable of objective answers i.e. avoid questions of opinion and keep questions of fact. For example, instead of asking a worker, "whether he is content with his present job, ask him if he desire to change his job and, if so, to what sort of job would he like to shift?

9. The number of questions should be kept to minimum. Unimportant questions having no bearing on the problem must be avoided. The precise number of questions would depend on the object and scope of investigation. Fifteen to twenty-five may be regarded as a fair number.

10. Arrangement of questions should be carefully planned. That is, the questions must be arranged in a logical order so that a natural and spontaneous reply to each is induced. As Croxton, Cowden and Klein have remarked. "Questions should not slip back and forth from one topic to another. For instance, it is illogical to ask a man how many children he has before asking whether he is married or not".

11. The questionnaire should provide necessary instructions to the informants. The instructions given must be precise and definite. All terms and units of measurement are to be clearly defined. For instance, if there is a question on weight it should be specified as to whether weight is to be expressed in pounds or kilograms. If necessary, examples of how to fill in the questionnaire may be given.

12. In respect of matters which are basic to the enquiry one or more cross checks should be incorporated into the questionnaire to verify whether the respondent is giving correct answers or not. For instance, in a family planning survey, one of the questions put to women is about their age. Two questions may be included in the questionnaire one may be 'what is your age' and another 'what is your date of birth?'.
3. In part A of the questionnaire we use the terms city town

13. The questionnaire should be pre-tested with a group before mailing it out. This will enable the investigator to spot out the short-comings in the questionnaire and revise it in the light of the tryout (i.e. trial or pilot survey)

We have given below a model questionnaire and a model schedule.

1. Model Questionnaire

Enquiry into The Expenditure Habits of Students Residing in College Hostels in The State of Tamil Nadu

Dear Sir/Madam,

It has been proposed to conduct a survey on the expenditure habits of students residing in college hostels in the State of Tamilnadu. The task of collection of data from colleges located within the Madurai District area has been entrusted to the Planning Forum of our college. Kindly answer the questions contained in the enclosed questionnaire and return it by post within a fortnight. A self-addressed stamped envelope is attached herewith for the purpose.

We assure you that the information supplied by you will be treated as confidential.

Thanking You,

Yours sincerely,

(Sd)

Secretary,
Planning Forum,
_____ College, Madurai.

Note:

1. By College hostel we mean hostel run by the college authorities. It may be run in the buildings owned by the college authorities or rented buildings.
2. In part C and part D of questionnaire you are asked to give figures for the different sources from which you are getting your income and for the different items on which you spend money. Kindly give MONTHLY figures or average monthly figures.
3. In part A of the questionnaire we use the terms city, town and village. If the population in a place is less than 10,000 it is to be treated as a village. If the population is 10,000 or more but less than one lakh, it is to be treated as a town. If the population is one lakh or more than one lakh it is to be treated as a city.

Questionnaire

Space for hints

A

- I. Name of the college _____
- II. Place of its location _____
- III. Is the place a city or town or village? _____
- IV. Is the college meant exclusively for men, women or for both _____

B

- V. Name of the student _____
- VI. Sex _____
- VII. Class studied _____
- VIII. Age with date of birth _____
- IX. Name of the Hostel in which the student is residing _____

C

1. Money received from parents/guardian in a month _____
2. Do you get any scholarship? _____
3. If so, the amount of scholarship got by you per month _____
4. Other sources from which you get money in a month:
 - a) From relatives _____
 - b) From friends _____
 - c) By having part time employment _____
5. Total money you get in a month _____

D

6. Money spent on account of tuition fee and other fees payable to the college in a month _____
7. Expenditure on text books, notes and note books in a month _____
8. Hostel food expenses in a month _____
9. Other hostel fees per month _____
10. Expenditure on clothing in a month _____

11. Expenditure on washing in a month _____
12. Expenditure on the purchase of toilet goods in a month _____
13. Expenditure on entertainment in a month _____
14. Are you in the habit of smoking? _____
15. If so, monthly expenditure on it _____
16. Expenditure on account of travel in a month _____
17. Expenditure on account of medical treatment in a month _____
18. Any other item or items for expenditure with money spent per month on each item _____
19. Total monthly expenditure _____

II. Model Schedule

Survey of Begging in Madurai City

A

1. Name of the beggar _____
2. Sex _____
3. Age _____
4. Religion _____
5. Place, District and state to which he/she belongs _____
6. Wanderer or permanently settled at Madurai _____
7. Place of residence at Madurai
 - a) House _____
 - b) Pavements and other places _____

B

8. Marital Status
 - (a) Single _____
 - (b) Married _____
 - (c) Widowed _____
 - (d) Divorced _____
 - (e) Separated _____

9. Defendants, if any

(a) Husband/wife _____

(b) Parents _____

(c) Children _____

(d) Other relatives _____

10. Number of defendants _____

11. Occupation of dependents

(a) Begging _____

(b) Studying _____

(c) Gainfully employed _____

(d) Remaining idle _____

C

12. Place of Begging

(a) Houses _____

(b) Shops _____

(c) Places of worship _____

(d) Bus Stand _____

(e) Railway Station _____

(f) Streets _____

13. Nature of aims got

a) Cooked food _____ b) Money _____ c) Other things _____

14. If money got through begging how much money got daily _____

15. How is this money spent _____

16. Saving, if any? _____

17. Is there any other source of income? _____

18. If so, the sources and amount got from each source _____

Space for hints

D

19. Is he/she a beggar by profession or by compulsion? _____
20. If beggar by profession, begging due to _____
- a) inheritance from parents
 - b) unwillingness to work
21. If beggar by compulsion begging due to _____
- (a) Orphan
 - (b) Physically and mentally unfit for work
 - (c) Insufficient income and debts
 - (d) Family tension and troubles
 - (e) Unemployment
22. Technique of begging _____
- (a) Plain appeal
 - (b) Exhibiting physical defects
 - (c) Street singing and dancing
 - (d) Carrying deformed persons as exhibits
 - (e) Religious orders
 - (f) Under religious disguise
 - (g) Cheating as physically unfit
23. Is the beggar literate or illiterate? _____
24. If literate, whether _____
- (a) able to read and write only
 - (b) completed primary education
 - (c) completed secondary education
 - (d) attended college

E

25. Whether willing to beg throughout the day _____
26. If not, how is the rest of the day spent _____
27. Whether willing to get employed in gainful occupation _____
28. If not, willing to get admitted in orphanage or beggar house _____

3.6 Secondary Data

Space for hints

(i) Meaning:

Secondary data are those which have already been collected, tabulated and presented in some form by someone for some purpose. In other words, secondary data are those which have gone through the statistical machine at least once. Secondary data, as its name itself conveys, is second hand data. If we take primary data to resemble 'raw materials' then we have to take secondary data to resemble 'finished products'.

(ii) Compilation and not collection:

In statistical investigation, the term collection is used to denote the assembling of entirely new data. Hence, strictly speaking, it is not correct usage to talk of collection of secondary data. The correct usage is COMPILATION of secondary data.

(iii) Sources:

The various sources of secondary data may be divided into two broad categories as noted below:

(a) Published Data:

Sources:

1) Official publications.

2) Semi-official publications.

3) Private publications.

Official publications include (a) reports and official publications of the Government and (b) reports and official publications of international bodies such as U.N.O., I.M.F., I.B.R.D., I.F.C., etc.

Semi-official publications include reports and publications of Municipalities, Corporations, District Boards, Life Insurance Corporation of India, the Unit Trust of India the Reserve Bank of India, Port trust etc.

Private publications include

(a) Publication of trade and professional bodies such as the Federations of Indian Chambers of Commerce, Institute of Chartered Accountants etc.

b) Publications of journals and newspapers such as 'Commerce', 'Capital', 'The Economic Times' etc.

c) Annual and periodical reports of the joint stock companies.

d) Publications of research bureaus, research scholars etc.

Types :

1) Continuous or Regular Data:

Statistical data published at short regular intervals are called continuous or regular data. The publication of weekly index number of wholesale prices, the publication of monthly figures of exports and imports etc., are examples of continuous or regular data.

2) Periodical Data:

Statistical data published at long regular intervals are called Periodical data, Indian census, Agricultural Statistics of India, Statistical tables relating to Banks in India etc., are examples of periodical data.

3) Irregular Data:

Certain types of data emerge from special studies connected with certain aspects of economic and social phenomena. They are irregular in nature that is, they do not have regular dates of publications. Examples are the reports of the National Income Committee, Tariff Commission reports and the reports of various committees and commissions appointed by the Government from time to time.

(b) Unpublished Data:

All Statistical materials are not always published. There are various sources of unpublished data. These include records maintained by various Government and private offices, studies made by research bureaus, labour bureaus, trade association, chamber of commerce, research workers etc., Such sources can also be used wherever necessary.

iv) Precautions in the use of secondary data:

The secondary data must be used with caution. In the words of W.I.King "When the investigation is to be of secondary type, it is necessary to exercise considerable care in several respects, before making use of the figures gathered by others". Scrutiny of the secondary data is essential because the data might be inaccurate, unsuitable, or inadequate. In the words of Bowley, "It is ever safe to take published statistics at their face value without knowing their meanings and limitations and it is always necessary to criticise arguments that can be based on them"

Before making use of the secondary data the investigator should follow the steps given below:

(a) Testing the Reliability of the secondary data :

Before using secondary data, one must get answers the following questions.

- 1) Who collected the data and from which sources?
- 2) Were the data collected by the use of proper methods?
- 3) Whether census method or sampling method was used in collection of data?
- 4) Whether the one collected them and source, both are dependable?
- 5) What degree of accuracy was desired by one who collected?
- 6) At what time were the data collected? Can it be regarded as normal time?
- 7) What was the purpose for which the data were originally collected?

(b) Ensuring suitability of the data

Even if the data are reliable, they should not be used if they are found to be unsuitable for the purpose of investigation. Data which are suitable for one enquiry may be entirely unsuitable for another. Suitability of the data can be checked by finding out

- 1) What was the object of the enquiry?
- 2) What was the standard of accuracy aimed at?
- 3) Do the data refer to homogeneous conditions?
- 4) Time of the collection of data.
- 5) The definitions of various terms and units of collection?

(c) Finding out adequacy of data:

Even if the data are reliable and suitable, it must be found out whether they are adequate or not. For instance, the data collected earlier may refer to an area which is narrower or wider than they are required for the present enquiry; if it is so, the data should be considered to be inadequate. Further the data may not cover suitable periods. For instance, the data collected earlier may contain yearly figures. But the present enquiry may be a monthly study of the phenomenon, in which, the data should be considered to be inadequate. Again, the degree of accuracy of the original data may be found to be inadequate for the present enquiry.

Thus, great care is necessary in arriving at conclusions when secondary data used owing to their limitations and inaccuracies that may be present. Hence, as Croxton, Cowden and Klein have remarked, "It is preferable to make use of primary source whenever possible".

4. SAMPLING

As we have already stated, there are two important techniques of data collection. They are:

- i) Census or complete enumeration technique and
- ii) Sample technique

4.1 Meaning of census and sample methods :

Suppose we want to test the strength of beams supplied by a firm. In order to test the strength, suppose we load each and every beam till it breaks. Now the technique adopted here is called census technique. But we cannot afford to adopt this technique in practice. Normally, we select only a few beams, test their strength and satisfy ourselves regarding the strength of all the beams supplied by the firm. The selected few beams are considered to be representative for the whole lot. The technique adopted here is called sample technique. Thus, under census method, each and every unit in which we are interested is taken into account whereas under sample method only few selected units are taken into consideration.

4.2 Population or Universe - Meaning:

In the above example, the set of all the beams supplied by the firm is called statistical population or universe. Thus, a statistical population or universe may be defined as the complete set of items belonging to our field of enquiry.

4.3 Types of Universe:

(i) Finite Universe

If the total number of items is a known number, the universe is a finite one. Total number of beams supplied by a firm is a known number and hence the universe is finite in this example.

(ii) Infinite Universe

Suppose we consider the universe of stars in the sky. Total number of stars in the sky cannot be determined and hence, is not a known number. In this case, the universe is called infinite universe.

Check your Progress

9. Define sample?

10. What is meant by population?

4.4 Sample - Meaning:

Space for hints

The selected few items which are considered to be representative of the population or universe constitute a sample. Thus, **sample is only a fraction of the population.**

The process of selecting a sample is called 'sampling'.

4.5 Sampling used by All:

In our every day life, we make use of the technique of sampling either consciously or unconsciously. When we go to a shop, we examine a handful of rice to find the quality of rice in the whole bag. The house-wife examines a ladle of rice to know if the rice in the pot is well cooked. A doctor examines a few drops of blood to draw conclusion about the blood constitution of the whole body. The teacher asks a student a few questions to determine his grasp of entire subject. A businessman places orders for a commodity by examining only a few units of the same. Thus, there is hardly any field in which the sample technique is not used.

4.6 Reasons for using Sample Method:

(i) Practicability:

The census method is impracticable in cases where destructive tests have to be applied or where the population is so large that it is physically impossible to consider every item in the population. For example, if we want to test the number of burning hours of 1 ...lbs supplied by a factory we cannot burn out all the bulbs to see if they are of proper specification. Similarly~ to test the quality of wheat in the bag, we cannot test the quality of each and every wheat in the bag. In the above cases, only the sample method is a practicable one.

(ii) Speed:

Under the sample method data may be collected and summarized much more quickly than under census method. This may be the most important consideration when the information is urgently needed or when it is to be used as a basis for government policy.

(iii) Accuracy:

More accurate results may be obtained from a sample enquiry than from census enquiry. The sample enquiry requires only a smaller staff and consequently better qualified staff can be obtained and they can be given better training. It may also be possible to spend more time and take greater care over each individual

questionnaire and to ask a greater number of questions. Thus, more detailed information may be obtained under sample method. Even though there may be errors in the sampling process, they are much less than the errors due to the inaccuracy and incompleteness of the large scale survey. Also, the errors due to sampling can always be determined and it is possible to devise ways of reducing the errors.

(iv) Cost:

Under sample method, as data are secured only from a portion of the population, expenditure will be much smaller than under the census method. This is a great advantage, particularly in an under developed economy where much of information would be difficult to collect by the census method for lack of adequate resources.

4.7 Disadvantages of Sample Method :

(i) Sample method is not applicable where information about every item of the population is required. To find out the total number of adults etc., sample method cannot be used. In such cases we have to resort to census method only.

(ii) Faulty methods of selection would make the sample not to be true representative of the population and the whole effort becomes a waste.

(iii) Sampling generally requires the services of experts for consultation purposes. In the absence of qualified and experienced persons, information obtained by the sample method cannot be relied upon.

(iv) Sampling techniques are based on theory of probability. Therefore, only those who are well versed in the theory of probability can use these methods.

4.8 Principles of Sampling :

The whole theory of sampling is based upon two important principles

(i) Law of statistical regularity, and

(ii) Law of inertia of large numbers

Law of Statistical Regularity :

Everything in nature and life occurs with a surprising regularity. Even accidents which are comparatively rare events do not happen haphazardly as people think. If 50 students are selected at random from a class containing 400 students, the average weight measurement for them will differ but little from

that of the entire class. If 100 sugarcane are taken at random from a field where sugarcane is grown and the average height of this sample is calculated it will not differ by more than a small fraction from the average height of all the sugarcanes in the entire field. These are manifestations of the law of Statistical Regularity. The law of statistical regularity is the explanation for the fact that a sample duplicates an entire population in all its characteristics.

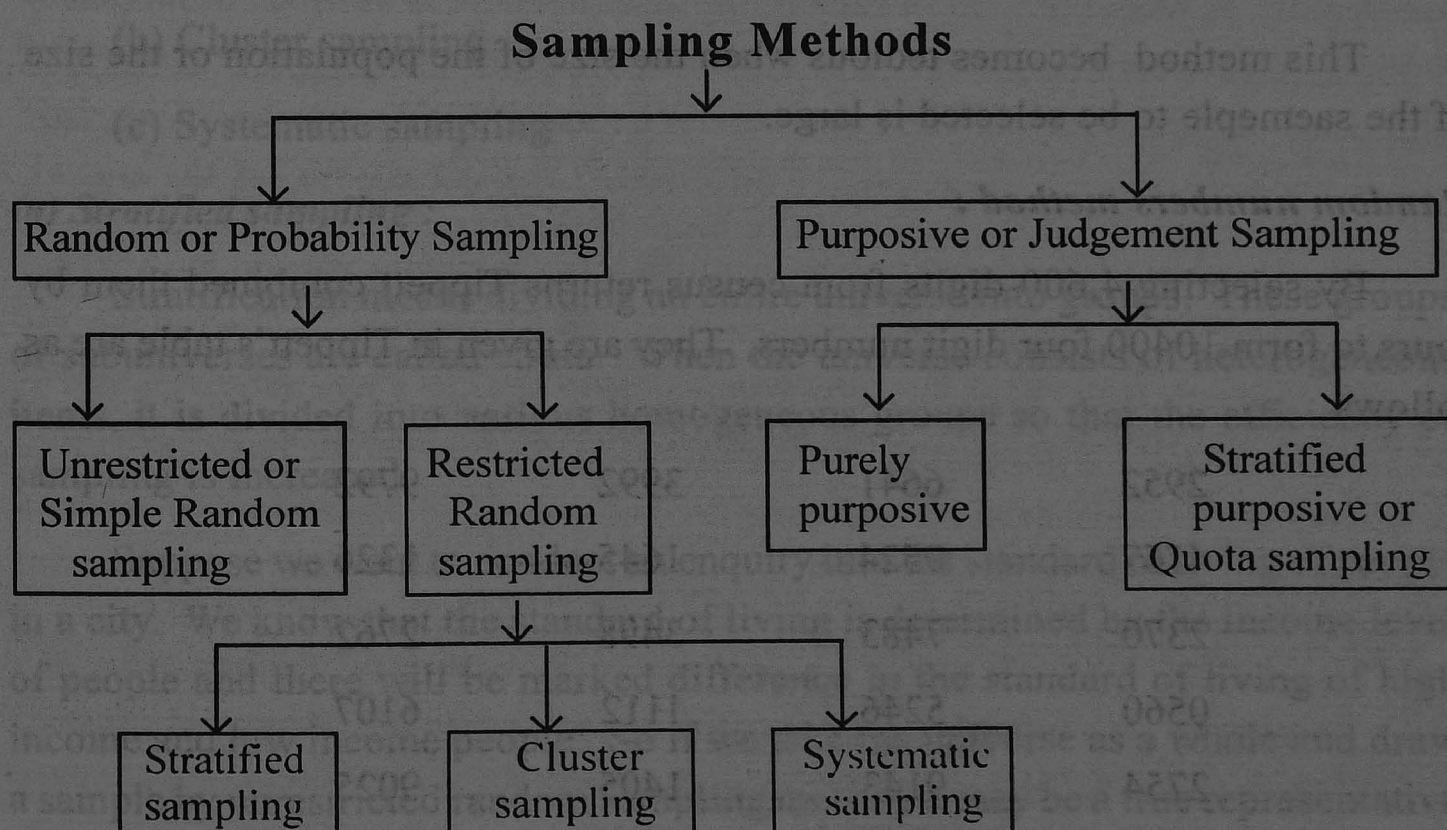
Law of Inertia of Large Numbers :

A natural consequence of the law of statistical regularity is the 'Law of inertia of Large Numbers'. It may be that in one year some have become paupers while a few have become rich in a particular city. Thus, there may be changes in individual items. But there would be very little change in the sum total of the wealth of the entire city. Thus there may be changes in individual items. But when we study the entire lot together, the change would not be appreciable. This is because the changes are not cumulative but compensating. That is while some items move in one direction others move in exactly opposite direction. This is said to be inertia of Large Numbers. This meaning is that large numbers together change slowly.

4.9 Methods of Sampling:

There are different methods of selection of samples. The choice of the suitable method depends upon (i) the nature of the problem under consideration (ii) the uses for which the results obtained from the sample are to be put.

Various methods of designing samples are presented below in the form of a tree diagram, so that you get a comprehensive picture of various sampling techniques available.



(A) Random Sampling :

Random sample is one where the individual items constituting the sample are selected at random "Random" is not used here in the sense of "haphazard" or "hit-or miss". By random selection we mean that the selection has been made in such a manner that each item of the population has equal or known opportunity of being included in the sample. Chance and only chance determines which items are to be included in the sample chosen.

(i) Types of Random sample

Random samples may be classified into two types; (a) unrestricted or simple random samples and (b) restricted random samples.

(ii) Methods of getting simple random sample:

Unrestricted random samples are usually obtained by using any one of the following two methods.

(a) Lottery method.

(b) Random numbers method.

Lottery Method :

The process of drawing by lot is also known as simple random sampling. In this method various item are represented by small chits of paper of the same size which are folded in the same manner and mixed together. From this the required numbers are picked out by a blind folded person. For example, if we want a sample having 10 items, then 10 chits are picked out from the required sample.

This method becomes tedious when the size of the population or the size of the saemeple to be selected is large.

Random numbers method :

By selecting 4,600 digits from census returns Tippett combined them by fours to form 10400 four digit numbers, They are given in Tippett's table are as follows.

2952	6641	3992	9792
4167	9524	1645	1326
2370	7483	3498	2762
0560	5246	1112	6107
2754	9143	1405	9025

Check your Progress

11. What is Random sampling?

12. What are the types of Random sampling?

The technique of selecting random sample with the help of these numbers is very simple. Suppose we have to select a sample of 15 items from a population of 4000 items. First we shall number the various items of population from 1 to 4000. We shall open any page of the Tippett's table and select the first 15 numbers which are less than 4000. The items having these numbers form the required random sample.

Through Tippett's number are of four digits, in many cases they are less than 1000 or 100 or 0. Consider the Tippett's numbers 0003, 0560, 0104. The number 0008 is simply 8 and it is less than 10. The other two numbers are less than 1006. Hence even if the population is small, say of 150 units only, and if we have to take a sample of 150 units only, and if we have to take a sample of 10, we can use Tippett numbers and get the sample.

Tippett numbers have been tested several times with known populations and the randomness of these numbers is almost beyond question. Similar tables have been devised by Fisher Kendall and Mahalanobis. They are also very popular and have given correct results in large number of investigations.

(iii) Methods of getting restricted random samples

So far we have seen the methods of getting unrestricted random samples, Unrestricted random samples usually result in greater cost or less degree of precision. Hence, in order to increase efficiency a restricted random sample is used, A restricted random sample may be obtained in any one of the following three methods:

(a) Stratified sampling

(b) Cluster sampling

(c) Systematic sampling

(a) Stratified sampling :

Stratification means dividing an entire universe into groups, These groups of subuniverses are called strata. When the universe consists of heterogeneous items, it is divided into various homogeneous groups so that the efficiency of sampling is increased.

Suppose we want to conduct an enquiry into the standard of living of people in a city. We know that the standard of living is determined by the income level of people and there will be marked difference in the standard of living of high income and low income people. So if we take the universe as a whole and draw a sample by unrestricted random sampling method it may be a true representative

of the universe. More reliable results may be obtained if we divide the population into high, medium and low income groups and then draw a sample.

The individual problem under investigation determines the characteristics that should be used for stratification. In practice geographical, sociological and economic characteristics are often used for the grouping of universe.

After the stratification is over, the next problem is to decide upon the number of items to be drawn from each stratum. We may draw equal number of items from each stratum or may draw terms in proportion to the size of each stratum. The latter case is called 'proportionate sampling', and the former, 'disproportionate sampling'. There is called another important method of representing each stratum known as 'optimum allocation method'. In this method, the variation of items within each stratum is also taken into account. Each stratum is represented in the sample according to the size and its variability.

After the method of representing each stratum has been decided upon, items are selected from each stratum by unrestricted random sampling method to get our required by unrestricted random sampling method to get our required stratified sample. Stratified sampling gives more accurate results compared to unrestricted random sampling.

(b) Cluster sampling :

As in the case of stratified sampling, here also the universe is divided into groups. But the method of dividing the universe into groups in cluster sampling is just opposite to that in stratified sampling. In stratified sampling is just opposite to that in stratified sampling the groups are such that (i) there is as much homogeneity as possible within each group and (ii) there is as much homogeneity as possible between the groups. But in cluster sampling, the universe is divided into groups or clusters such that (i) there is as much heterogeneity as possible within each cluster and (ii) as small a difference as possible between the cluster.

Suppose we want to conduct an enquiry into the wages earned by workers in an industry. The industry may consist of several companies. The industry is our universe and is divided into groups each group being a company. Unlike in the case of stratified sampling, in cluster sampling, we select few clusters or groups by unrestricted random sampling method. In our example few companies are selected by unrestricted random sampling method. The clusters (i.e. companies, in our example) are called primary sampling units.

There may be a number of employees in each company. These employees are called elementary sampling units. Now, can either investigate each and every

elementary sampling unit in each selected cluster or we can draw an unrestricted random sample or elementary sampling units from each selected cluster. That is we can take into account each and every employee in each selected company or we can draw a sample of employees from each selected company by unrestricted random sampling method.

Clusters formed on a geographical basis are of great practical importance. Suppose we want to study grocery store sales in Tamil Nadu. The elementary sampling unit is grocery store. The Tamil Nadu State may be divided into a number of districts and few districts may be selected at random. Thus, the districts form the Primary sampling units. Random sample of grocery stores from each selected district may be drawn and studied.

As a general rule, a cluster sample is considerably less expensive than an unrestricted random sample of equal size. It saves time also. In the above illustration the interviewer has to travel between the locations of elementary sampling units (i.e., grocery stores). Considerable amount of money and time will be saved if cluster sampling is used rather than unrestricted random sampling method.

Furthermore, lack of complete and upto date list of elementary sampling units makes it impossible to use unrestricted random sampling method. In such cases cluster sampling will be the only practical solution.

There is, however, an important limitation of cluster sampling method (i.e.) this method generally gives less accurate results compared to other methods of sampling.

(c) Systemetic sampling :

While stratification and cluster sampling involve the division of the universe into groups, systemetic sampling involves an ordering of the universe. The ordering may be alphabetical, numerical, geographical or any other.

After the items in the universe have been ordered, every 10th or 20th or 30th items are selected depending upon the size of the universe and the size of the sample to be selected.

For example, suppose there are 150 rooms in a hostel each with a serial number and we have to select a sample of 10 students out of the 150 students residing in these rooms. Here we can select the students in every 15th room. The students may be arranged in another way also the name of the students may

be written in alphabetical order. From this arrangement also; every 15th student may be selected and thus we can get the sample. The number 15 is obtained by dividing the number of items in the universe viz, 10. This number 15 is called sampling interval.

Now the the text question is, from where to start the sample selection. We may start sample selection anywhere in between the first and 15th student in the arrangement. A random start is always preferable. Hence, a number is selected at random between 1 and 15. Suppose the number selected is 6. Now, we start selection from the 6th student. This is, the students corresponding to the numbers 6, 21, 36, 51..... are selected to complete the sample. Systematic sample will be a random sample for all practical purposes. Hence, it is also known as quasi-random sample.

The main advantage of systematic sampling is its simplicity. An advantage of stratification can also be realized by an appropriate ordering of the uiverse.

However, systematic sampling should not be used under the situations which are similar to the situation explained below :

Suppose, we have to select a sample of households in a block. Also suppose that every eighth house in the block happens to be a big house at the corner. Here, if the sampling interval also happens to be eight, then the systematic sampling method would yield a sample of only big corner houses. This sample will not serve our purpose.

Thus, if the items in the population at the end of each of a particular interval are alike systematic sampling method should not be adopted.

Upto this point our discussions on sampling have referred to random or probability sampling. We now turn to an other method of sampling known as purposive sampling.

B. Purposive or Deliberate or Judgement Sampling :

Purposive sampling means selecting the items in accordance with some purposive principle or in accordance with someones personal judgement. Here, the chance of the inclusion of some item of the population in the sample is very high while that of others is very low.

While choosing the purposive sample, only the average items are considered and extreme items are omitted. In accordance with the object of the enquiry, the selection of the sample is adjusted so that no significant item may be ignored.

It is clear that in this method the bias of the investigator can play a very

important role and destroy the representativeness of the sample. But there may be cases where a purposive sample may give better results than a random sample. For example, if we have to select only 5 students out of 5000 with a view to study their height measurements purposive selection can give better results than random selection. The reason is that the investigator would select such five students who have normal or mean height according to his judgement. But however, if the size of the sample is larger, random sample would give better results than purposive sample. This is due to the law of inertia of large numbers.

C. Quota sampling :

It is a type of purposive sample. Quota sample, is basically a stratified purposive sample: In a quota sample, the total sample is broken down on the basis of known strata in the universe (for instance) age, sex, occupation, social class etc.) Then quotas (that is number of people to be interviewed from each group) are fixed according to the proportion of people belonging to different strata in the universe leaving the selection of actual respondents to the complete discretion of the interviewers. For example, in a radio listening survey, the interviewers may be asked to interview 100 persons living in a locality and that out of these 100 persons, 50 must be housewives, 30 must be farmers and 20 must be children below the age of 15. Within these quotas fixed, the interviewer is free to select the persons to be interviewed.

In this method, the cost per person interviewed will be very small. But considerable amount of personal bias is likely to be present. For example, the interviewer may miss farmers working in the fields and talk with those housewives who are at home. If a person refuses to respond, the interviewer may simply select some other person.

Quota sampling method is very popular in market surveys and public opinion polls.

Comparison of probability and purposive sampling :

The whole sampling theory is based on chance selection and hence its principles are not applicable to purposive sampling. Consequently, the reliability of the conclusions drawn from a purposive sample about a characteristic of the universe cannot be established by sampling theory.

If the size of the required sample is small, purposive samples give better results than probability samples. But in the case of probability sampling, increase in the size of the sample leads to increase in the precision of estimates and this may not be true in the case of purposive sampling.

Cost and time are important factor in practical sampling work. Purposive sampling often requires less money outlay and less total time compared to probability sampling.

In pilot studies often purposive samples are used by the statisticians.

5. Frequency Distribution

5.1 Introduction :

In any statistical enquiry, we collect data, arrange them, analyse them and Interpret them. The data collected may be used as they are and be analysed. In that case, we are said to analyse ungrouped data. Instead, the data collected may be grouped in some form and be analysed in which case we are said to analyse grouped data. Thus the data used in analysis may be either UNGROUPED OR GROUPED.

The data may be grouped in two ways. They are

- 1) Discrete frequency distribution method
- 2) Continuous frequency distribution method.

An example will enable the students to understand the difference among these three forms of data. Suppose the incomes of 11 persons are as given below:

Rs.100, Rs.100, Rs.125, Rs.125, Rs.125, Rs.160, Rs.180, Rs.200, Rs.200, Rs.200, Rs.200.

These data are called ungrouped data.

Suppose we say that two persons get an income of Rs.100; three persons get an income of Rs.125; one person gets an income of Rs.160; one person gets an income of Rs.180 and four other persons get an income of Rs.200. Now we have grouped the given data by finding out the number of times each particular value occurs. We can present this in the form of a table as given below:

Size of Income (Rs.)	Number of persons getting it
100	2
125	3
160	1
180	1
200	4

Check your Progress

13. Distinguish between grouped and ungrouped data?

The above table is called by the name of “Discrete Frequency distribution” .

Space for hints

Suppose instead of counting the number of persons who receive a particular size of income. We fix various income ranges and try to find out number of persons who come within each range. Suppose the income ranges fixed are Rs.100 to Rs.150 and Rs.151 to Rs.200. In our illustration, incomes of five persons fall within the range of Rs.100 to 150; incomes of six persons fall within the range the range of Rs.151 to Rs.200.

Now we have grouped the given data by finding out the number of values falling within each income range which we have chosen. We can present the data grouped in this way in the form of a table given below :

Income Range	Number of persons
Rs. 100 to Rs. 150	5
Rs.151 to Rs. 200	6

The above table is called by the name of ‘**Continuous Frequency Distribution**’.

There are various steps involved in grouping the **raw data** (i.e.ungrouped data) into discrete frequency distribution and continuous frequency distribution. In grouping them, we make use of certain technical terms. In this topic, we have explained in detail the various steps involved and the meanings of these technical terms. The students are advised to become familiar with these steps and terms. This will enable them to understand the concepts and to do problems coming in the succeeding lessons.

5.2 Ungrouped Data:

The data we gather in a statistical enquiry are called raw data or ungrouped data. Suppose the following figures are the marks obtained by 45 students of B.Com at a University Examination.

68, 90, 75, 74, 63, 86, 61, 62, 84, 62, 72, 77, 72, 65, 73, 79, 75, 88, 60, 85, 81, 72, 68, 88, 82, 68, 68, 75, 72, 90, 68, 75, 68, 90, 75, 76, 68, 66, 76, 80, 68, 68, 68, 68, 66.

5.3 Item and Array:

The above figures are raw data or ungrouped data. Each of the above figures is called an item. In the raw data, the items are in a haphazard order of size and hence only a very little information is forthcoming. It would be a tedious task to find even the lowest and highest marks. So, the items are arranged in ascending or in descending order of magnitude to facilitate the analysis of data. An arrangement of items in ascending or in descending order of magnitude is called an array. For example, let us arrange the marks obtained by the 45 students of B.Com in ascending order of magnitude as follows:

60, 61, 62, 62, 63, 65, 66, 68, 68, 68, 68, 68, 68, 68, 68, 68, 68, 68, 68, 72, 72, 73, 73, 74, 75, 75, 75, 75, 75, 75, 76, 77, 78, 79, 80, 81, 82, 84, 85, 86, 88, 88, 90, 90, 90.

This is called an array. The array enables us to see at once the maximum and minimum values in the given data and gives a rough idea of the distribution of items.

The array, however, is a cumbersome form of the data. When the number of items to be arranged is quite large, the process of arranging becomes tedious. Also, on several occasions we are unable to form sharp and clear cut impression about data from the array formed. Hence, condensation of the array into more usable form is essential.

5.4 Frequency:

In the data collected, some values occur only once whereas other values occur more than once. In the example we have given earlier, it is to be noted that the value 68 occurs twelve times, 75 occurs six times; 90 occurs thrice; 62, 72, 73 and 88 occur twice; and all the other values occur only once.

Number of times a value occurs is called frequency of that value. In the above example, 12 is the frequency of the value 68; 6 is the frequency of the value 75; 3 is the frequency of the value 90; 2 is the frequency of each of the values 62, 72, 73 and 88; and 1 is the frequency of all the other values.

5.5 Discrete Frequency Distribution:

Now the table formed with the two columns viz., value of item and frequency is called discrete or discontinuous frequency distribution or simply frequency array.

Discrete frequency distribution is formed as follows.

First, under the heading 'value of item' the first value in the given set of

data is written. The occurrence of this value is marked by a small vertical line (called tally mark) drawn against it under the heading 'tally marks'. Now we look for the second value in the data. If it also happens to be the same value as in the first place, then another tally mark is made against the first value by the side of the first tally mark. If the second value happens to be different from the first value, it is written below the first value under the heading 'value of item' and a tally mark is made against it. Like this all values in the given set of data are written under the heading 'value of item' and the occurrence of a value for each time is marked by a tally mark made against it, under the heading 'Tally marks'. So the number of tallymarks made against each value gives the number of times each value occurs. i.e., it gives the frequency of each value. Therefore, we count the number of tallymarks against each value and give those numbers against each value under the heading 'frequency'.

If a value occurs a large number of times, then it would be a tedious task to count the number of tally marks marked against it. Therefore, for convenience in counting the tally marks, we adopt the following procedure in marking the tally marks.

Suppose a value occurs five or more than five times. The occurrence of that value for the first four times is marked by drawing four small vertical lines side by side. The occurrence of the value for the fifth time is marked by a small horizontal line drawn over the previous four vertical lines. Now, the occurrence of the value for the sixth time is marked by small horizontal line drawn over the second four vertical lines. Now we have two blocks. If the same value occurs for the eleventh time also, a small vertical line is drawn again by the side of the second block and so on. Now it is easier to count the number of tally marks. Because, it is enough if we count the number of blocks. and. number of vertical lines by the side of the last block; number of blocks multiplied by five is added to the number of tally marks by the side of the last block to give the total number of tally marks.

Let us consider the same example of marks obtained by 45 students of B.Com., class at a University Examination. In this example, 68 occurs twelve times. The occurrence of 68 for the first four times are marked by four small vertical lines and its occurrence for the fifth time is marked by a horizontal line drawn over them. Its occurrence for next five times is marked by another block of four vertical lines and one horizontal line and its occurrence for the 11th and 12th times are marked by another two vertical lines by the side of the second block (see the table given below). Now, we have two blocks. Two multiplied by five is ten and it is added to two which is the number of tally marks by the side of

the second block. We get the sum to be twelve and it is noted against 68 under the heading frequency. Like this we have formed the discrete frequency distribution of the marks obtained by the 45 students of RCom.class and given it below:

Value of item (Marks)	Tally -Marks	Frequency
60	I	1
61	I	1
62	I	2
63	I	1
65	I	1
66	I	1
68	IIII II	12
72	II	2
73	II	2
74	I	1
75	IIII I	6
76	I	1
77	I	1
78	I	1
79	I	1
80	I	1
81	I	1
82	I	1
84	I	1
85	I	1
86	I	1
88	II	2
90	III	3
Total		45

The above table clearly shows the number of students who scored various marks. We are easily able to find that maximum number of students have got 68 marks, 3 students have got the maximum marks and so on. But this sort of condensation is useful only when the values occur frequently. If it is not so, benefits from condensation of data cannot be reaped. In the above table it is to

be noted that a number of values occur only once; there is still here for further condensation. A better method is to divide the data into classes instead of giving the number of times each value occurs as in the case of discrete frequency distribution. This method of grouping is called continuous frequency distribution.

5.6 Continuous Frequency Distribution - Related concepts

Before giving the definition and construction of a continuous frequency distribution the following terms should be known to you.

(i) Class Interval:

Consider a set of data relating to the marks obtained by a set of students in an examination. We can divide this data into several groups such that the marks within 0 to 9 form one group. Within 10 to 19 form another group. Within 20 to 29 form another group and so on. Now 0 to 9, 10 to 19, 20 to 29, etc., are called 'class intervals'. These class intervals are written as 0-9, 10-19, 20-29, the hyphen mark representing 'to'.

(ii) Class Limits:

When we write a class interval, we write down two numbers which are known as 'class limits'. In our example, 0 is the lower class limit and 9 is the upper class limit of the class interval 0-9, 10 is the lower class limit and 19 is the upper class limit of the class interval 10-19 and so on.

(iii) Inclusive Method:

In the above example, not only the students whose marks are lying between 0 and 9, but also the students whose marks are either equal to 0 or equal to 9 are included in the class interval 0-9. In the case of the remaining class intervals also we include the students in each class interval in a similar manner. That is, the items which may be included in each class interval are:

(i) equal to the lower limit of the class interval

or

(ii) lying between the lower limit and upper limit of the class interval.

or

(iii) equal to the upper limit of the class interval.

Now this method of grouping is called 'inclusive method'.

(iv) Not true class intervals and true class intervals:

This inclusive method class intervals are also known as not true class intervals. In this type of class intervals, the upper limit of each class is not equal

to the lower' limit of the next class and there is a gap in between any two successive class intervals. For example, 20-24, 25-29, 30-34 and 35-39 are inclusive method class intervals or not true class intervals. Here the upper limit of the first class is 24 and the lower limit of the second class is 25 and the limits are different. The upper limit' of the second class (viz., 29) is different from the lower limit of the third class (viz., 30). The upper limit of the third class is different from the lower limit of the fourth class and so on.

This type of class intervals are useful only when the variable is a discrete one. Let us consider a discrete variable, say number of workers per factory. This variable will take only whole numbers as its values. We will never get a value like 24.3 or 29.7 workers, etc. So all the possible values of the given discrete variable will be included in the not-true class intervals even though there is a gap between two successive class intervals.

Suppose the given variable is a continuous variable. For example, let us take the variable, age. Age of a person need not be a whole number. The age of a particular person may be 24 years and 4 months. That is, the age is 24.3 years approximately. If we take the class intervals to be 20-24, 25-29 and 30-34, in no class interval we can include the value 24.3. Because, 24.3 is greater than the upper limit of the first class viz, 24 and hence we cannot include this item in the first class; similarly, 24.3 is less than the lower limit of the second class viz., 25 and hence it cannot be included in the second class also. So, the inclusive method class intervals cannot be used in the case of continuous variable.

In the case of continuous variable, we take the class intervals in such a way that there is no gap between any two class intervals. That is, we fix the class intervals in such a way that the upper limit of each class is equal to the lower limit of the next class. For example, we take the class intervals as 20-25, 25-30, 30-35, etc., This type of class intervals are known as 'True class Intervals'.

Here, another type of problem arises. In case the value of an item is exactly equal to the upper limit of a class interval, the problem is whether to include this item in that class itself or in the next class? We solve this problem by making the following assumption. In the case of true class intervals, if the value of an item is exactly equal to the upper limit of a class it is not included in that class but is included in the next class. In our example, if the age of a person is exactly equal to 25 years, he is not included in the class 20-25 but he is included in the class 25-30. If the age of a person is exactly equal to 30 years, he is not included in the class 25-30 but he is included in the class 30-35 and so on. So, in

the case of true class intervals, items which may be included in each class interval are,

Space for hints

(i) equal to the lower limit of the class interval

or

(ii) greater than the lower limit but less than upper limit of the class interval.

In the case of true class intervals, the class limits are known as 'true lower limit' and 'true upper limit'. The true lower limit and true upper limit are also called 'lower boundary' and 'upper boundary' of class interval.

(v) Exclusive Method:

When we have true class intervals, items whose values are exactly equal to the upper limit of each class interval are not included in that class interval; in other words, (excluded from that class interval) and are included in the next class interval. So, this method of grouping is known as 'exclusive method'.

(vi) Conversion of Not - True Class Intervals into True Class Intervals:

Generally, inclusive method is not in use as the element of continuity of the class limit is lost. In the calculation of certain measures in the analysis of data, if the class intervals are in the inclusive form they are to be transformed into the exclusive form. So we explain below the method of getting the true class intervals from the given not true class intervals.

1) First find out the difference between the upper limit of the first class and lower limit of the second class.

Let 20-24, 25-29 and 30-34 be the given class intervals. To convert these intervals into true class intervals, first of all, we must find out the difference between the upper limit of the first class viz., 24 and the lower limit of the second class viz., 25. The difference is $25 - 24 = 1$.

2) Next we have to divide the difference (which we have calculated above) by two. In our example, the difference is 1. When it is divided by 2 we get the number $\frac{1}{2} = 0.5$.

3) Now this 0.5 is to be subtracted from each of lower limits of the given class intervals. The resultant figures are the true lower limits of the respective class intervals.

The lower limits of the given class intervals are 20, 25 and 30. If we subtract 0.5 from each of these lower limits, we get the numbers 19.5 ($= 20 - 0.5$), 24.5 ($= 25 - 0.5$), and 29.5 ($= 30 - 0.5$). So, 19.5, 24.5, and 29.5 are the true lower limits of the first, second and third classes respectively.

4) Next, the number 0.5 is to be added to each of the upper limits of the given class intervals. The resultant figures are the true upper limits of the respective classes.

The upper limits of the given class intervals are 24, 29 and 34. If we add 0.5 to each of these upper limits, we get the numbers 24.5 ($= 24 + 0.5$), 29.5 ($= 29 + 0.5$), and 34.5 ($= 34 + 0.5$). Now, 24.5, 29.5 and 34.5 are the true upper limits of the first, second and third classes respectively.

So, in our example, the true class intervals are 19.5-24.5, 24.5-29.5 and 29.5-34.5 respectively.

Let us consider another example. Let 0-9, 10-19, 20-29, 30-39 be the given class intervals. We convert these intervals into true class intervals as follows.

The difference between the upper limit of the first class and lower limit of the second class is $10 - 9 = 1$.

When this difference is divided by 2 we get the number $1/2 = 0.5$.

If we subtract 0.5 from each of lower limits of the given class intervals, we get the numbers -0.5, 9.5, 19.5 and 29.5. These are the true lower limits of the given class intervals.

If we add 0.5 to each of upper limits of the given class intervals~ We get the numbers 9.5, 19.5, 29.5 and 39.5. These are the true upper limits of the, given class intervals.

Therefore, the true class intervals are -0.5 - 9.5, 9.5-19.5, 19.5-29.5 and 29.5-39.5.

In the table below we have given class intervals and the true class intervals.

Given Class intervals	True class intervals
0-9	0.5 - 9.5
10-19	9.5-19.5
20-29	19.5-29.5
30-39	29.5-39.5

(vii) Length of Class Interval:

The difference between the true upper limit and the true lower limit of a class interval is called the length or magnitude or Size or width of the class interval. Thus,

Length of a class = (True upper limit - True lower limit) of that class.

For example, when the class intervals are 19.5-24.5 and 24.5-29.5, the length of the first class is $24.5 - 19.5 = 5$ and the length of the second class is $29.5 - 24.5 = 5$. Here, we are given true class intervals directly. If we are not given true class intervals, to find out the length of the class, we must first convert the given class intervals into true class intervals. Suppose, 20-24 and 25-29 are the given class intervals. Here to find out the length of each class they are converted into true classes as 19.5-24.5 and 24.5-29.5. Now the difference between the class limits of each true class gives the length of each class.

If the length of each class is the same, then the class intervals are said to be 'equal class intervals'. For example the class intervals viz., 19.5-24.5, 24.5-29.5, and 29.5-34.5 are equal class intervals. On the other hand, if the length of each class is not the same then the class intervals are called 'unequal class intervals'. For example, the class intervals viz, 19.5-24.5, 24.5-30.5, 30.5-40.5 are unequal class intervals. Because, the length of the first class is 5, of the second class is 6 and of the third class is 10 and the three lengths are not the same.

The formula to calculate the length of the class intervals is as follows:

$$\text{Length of the class} = (\text{True upper limit} - \text{True lower limit})$$

We can also write this formula as follows:

$$(\text{True upper limit} - \text{True lower limit}) = \text{Length of the class.}$$

$$\text{That is, True upper limit} = \text{True lower limit} + \text{Length of the class.}$$

So using this formula we can get the true upper limit of a class interval if we are given the true lower limit and the length of the same class.

Let 70 be the true lower limit and 5 be the length of a particular class. Now, the true upper limit of this class is $70 + 5 = 75$.

Above we have seen that the true upper limit of a class can be got if we are given the true lower limit and the length of the same class. In the same way, using the following formula, we can get the true lower limit of a class if we are given the true, upper limit and the length of the same class.

$$\text{True lower limit} = (\text{True upper limit} - \text{Length of the class}).$$

Let 40 be the true upper limit and 5 be the length of a class. Now the true lower limit of this class is $40 - 5 = 35$.

(viii) Midvalue or class mark:

The value which lies midway between the lower and the upper limits of a

class is called the 'midvalue' or the 'class mark' of that class. The midvalue is calculated as follows:

We must first find out the sum of the lower and upper limits of the given class. This sum is to be divided by two. The resultant number is the midvalue of the given class. That is,

$$\text{Midvalue} = \frac{\text{Lower class limit} + \text{Upper class limit}}{2}$$

Suppose 19.5-24.5 and 24.5-29.5 are the given class intervals.

$$\text{The midvalue of the first class} = \frac{19.5+24.5}{2} = \frac{44}{2} = 22$$

$$\text{The midvalue of the second class} = \frac{24.5+29.5}{2} = \frac{54}{2} = 27$$

Suppose the given class intervals are 20-24 and 25-29

$$\text{The midvalue of the first class} = \frac{20+24}{2} = \frac{44}{2} = 22$$

$$\text{The midvalue of the second class} = \frac{25+29}{2} = \frac{54}{2} = 27$$

From the above example, it is clear that whether the given class intervals are true class intervals or not, the midvalues are the same. So, when we have to find out the mid values, we need not see whether the given class intervals are true class intervals or not.

It is to be noted that in the case of equal class intervals, the length of the class interval would also be equal

- (i) to the difference between the lower limits of any two successive classes;
- (ii) to the difference between the upper limits of any two successive classes and
- (iii) to the difference between the midvalues of any two successive classes.

For example, let 20-24, 25-29 and 30-34 be the given class intervals. We have already seen that the length of each of the above classes is equal to 5.

The difference between the lower limits of the first two classes is $25-20 = 5$. The difference between the lower limits of the second and third classes is $30-25 = 5$. Hence, the length of the class intervals is equal to the difference between the lower limits of any two successive classes.

The difference between the upper limits of the first two classes is $29-24 =$

5. The difference between the upper limits of the second and third classes is $34 - 29 = 5$. Hence, the length of the class intervals is equal to the difference between the upper limits of any two successive classes.

The midvalues of the given class intervals are 22, 27 and 32 respectively. The difference between the first two midvalues is $27 - 22 = 5$. The difference between the second two midvalues is $32 - 27 = 5$. So the length of the class intervals is equal to the difference between the mid values of any two successive classes.

Some times midvalues of the class intervals may be given and we may have to find out the class intervals from these midvalues. So, we have explained below the method of determination of class intervals from the midvalues:

- 1) First find out the difference between any two successive midvalues.
- 2) Divide the difference calculated above by 2.
- 3) The resultant figure is to be added to each midvalue to get the upper limits.
- 4) The resultant figure is to be subtracted from each midvalue to get the lower limits.

Let us consider the following example.

Midvalue
5
15
25
35
45

We find out the class intervals from the above midvalues as follows:

The difference between first two midvalues is $15 - 5 = 10$. If we divide this difference by 2, we get number 5. Now, we have to add this number 5 to each midvalue. The numbers we obtain by making this addition are 10 ($= 5 + 5$), 20 ($= 15 + 5$), 30 ($= 25 + 5$), 40 ($= 35 + 5$), 50 ($= 45 + 5$). So, 10, 20, 30, 40, 50 are the upper limits.

Next we have to subtract 5 from each midvalue. The numbers we obtain by making this subtraction are 0 ($= 5 - 5$), 10 ($= 15 - 5$), 20 ($= 25 - 5$), 30 ($= 35 - 5$), 40 ($= 45 - 5$). So, 0, 10, 20, 30, 40 are the lower limits.

Space for hints

The class intervals corresponding to the given midvalues are 0-10, 10-20, 20-30, 30-40 and 40-50.

We have given below midvalues and the class intervals in tabular form:

Class intervals	Midvalues
0-10	5
10-20	15
20-30	25
30-40	35
40-50	45

ix) Frequency:

The number of items which fall in an interval, is called the 'frequency' of the class. For example if 15 items have their values in the interval 19.5-24.5, then 15 is the frequency of the class.

Now, we give the definition of continuous frequency distribution as follows:

5. 7 Definition of Frequency Distribution:

When a set of values of a variable are grouped into several classes and these classes are arranged in the order of magnitude with the corresponding frequency against each class, the result is a 'frequency distribution' or 'frequency table'. It is also known as a simple frequency distribution.

When have given below an example of continuous frequency distribution.

Class Interval	Frequency
0 - 10	2
10 - 20	5
20 - 30	3
30 - 40	8
40 - 50	15
50 - 60	10
60 - 70	7
Total	50

Check your Progress

14. Define frequency distribution?

The frequency table may also be given in the following form.

Space for hints

Class interval	0-10	10-20	20-30	30-40	40-50	50-60	60-70	Total
Frequency	2	5	3	8	15	10	7	50

5.8 Construction of continuous frequency table:

There are two major steps involved in the construction of a continuous frequency table:

(a) Choice of class intervals and class limits

(b) Recording the frequency of each class.

(i) Choice of class intervals and class limits:

There is no hard-and-fast rule regarding the choice of the number of class intervals. If there are too many classes, advantages of summarization should be destroyed. If there are too few classes so many frequencies would be crowded into each class causing much information to be lost. The number of classes to be chosen depends partly upon the nature of the data and partly upon number of items in the given data. The greater the number of items, the more classes we may have. Also, if the difference between the highest and the lowest values in the given data is vast, we may have greater number of classes. In general, it might be stated that less than 5 classes and more than 70 classes should rarely be used. More than 20 or 25 classes would be useful only for working with extensive data.

Choice of the magnitude of the classes depends upon the range, (i.e., the difference between the highest and lowest values in the data) number of classes and the representativeness of the midvalues of classes. If the range is large the magnitude of the classes can also be large and vice-versa. But this should not make the number of classes too few or too many.

In most cases, it is necessary to represent the class intervals by single values. In such cases, middle value of each class is taken as the value of all the items belonging to that class. Hence, magnitude of the class should be small. Also, as far as possible the class limits should be so chosen that values of the given items do not coincide with either of the two limits of each class and atleast approximately coincide with the midvalues of class intervals.

Magnitude of the class intervals in multiples of 5 like 10, 20, 25 etc., must be preferred. As far as possible values of class magnitudes, like 3, 7, 11 etc., should be avoided. Equal class intervals are preferred to unequal class intervals.

There is no harm in using unequal class intervals if there is any good reason to do so.

The position or starting point of a class must either be 0 or 5 or multiples of 5. For instance, if the lowest value in the given data is 17 and if it is desired to have a class interval of magnitude 5, then the first class interval should be 15 to 20 and not 17 to 22.

(ii) Recording the frequency of each class:

After the fixation of class intervals, each item is considered one by one; a tally mark is made against the class to which it belongs; occurrence of items for fifth time in a class is marked by a small horizontal or slant line drawn over the previous four tally marks. Number of tally marks against each class is counted and recorded.

Let us form continuous frequency distribution of the data given below:

67, 34, 36, 48, 31, 61, 34, 43, 45, 38, 32, 27, 61, 29, 47, 36, 50, 46, 30, 46, 32, 46, 30, 33, 45, 49, 48, 41, 53, 36, 55, 64, 45, 25, 54, 58, 24, 33, 46, 45

The highest value = 67

The lowest value = 24

Let us have the class intervals 20-29, 30-39, and so on. The last class interval will be 60-69. Now the frequency table is given as follows:

Class Intervals	Tally marks	Frequency
20-29		4
30-39		13
40-49		14
50-59		5
60-69		4

5.9 Continuous frequency distribution with Open end Class :

On several occasions there is the possibility of having very few number of low or high values in the given data. For example, only one or a very few number of students may get one digit marks and all the other students may get above 40. In such a case, while forming the frequency distribution the first class is taken

‘below 40’ (or Less than 40). The lower limit of the first class is left unfixed; that is, the lower end of the first class is kept open. In this case, the frequency table will be of the following form:

Space for hints

Class intervals (Marks)	Frequency (No. of students)
Below 40	2
40-50	7
50-60	15
60-70	5
70-80	4

Here the first class is called an open class and the distribution is said to have open end at the beginning.

On the other hand, let us suppose that a few students got marks above 95 and other students got marks below 70. In this case, while forming the frequency distribution, the last class is taken as ‘Above 70’. Thus the upper limit of the last class is left unfixed. That is, the upper end of the class is kept open. In this case, the frequency table will be of the following form:

Class interval (Marks)	Frequency
30-40	4
40-50	8
50-60	7
60-70	5
above 70	3

Here the last class is called open class and the distribution is said to have open at its end.

There is another possibility also. A very few number of students may get one digit marks and another set of students may get marks above 95. All the other students may get marks between 40 and 70. In such cases, while forming the frequency distribution, the first class is taken as ‘Below 40’ and the last class is taken as ‘Above 70’. In this case, we have two open classes, one at the beginning and the other at the end of the distribution and the frequency table will be of the following form.

Class interval (Marks)	Frequency
Below 40	2
40-50	8
50-60	1
60-70	7
Above 70	1

Thus we can define open class interval as that class interval in which one of the two limits is left unfixed.

Also it is to be noted that the open class may come at the beginning or at the end or at the beginning as well as at the end of a frequency distribution. If the open class comes at the beginning of a distribution the lower limit of the first class is not known. If the open class comes at the end of a distribution, the upper limit of the last class is not known. If the open class comes both at the beginning and at the end of the distribution, the lower limit of the first class and the upper limit of the last class are not known.

The frequency distribution, containing an open class is called '**open end distribution**'.

While forming a frequency distribution, intervals having zero frequencies are avoided with the help of open class intervals.

5.10 Cumulative Frequency Distribution:

Sometimes we need answers to question like "how many students got marks less than 30?" or "how many students got marks less than 40?" etc. To answer these questions we derive a distribution called 'less than cumulative frequency distribution' from the given simple frequency distribution.

In some other cases, we may need answers to questions like 'how many students got marks more than 0?' or 'how many students got marks more than 40?' etc. Here also, to answer the above questions, we derive a distribution called 'more than cumulative frequency distribution' from the given simple frequency distribution.

(i) Derivation of Less than cumulative frequency distribution :

First of all, we give the definition of less than cumulative frequency' and less than cumulative frequency 'distribution'.

When we are given a simple frequency distribution, the number of items whose values are less than the upper boundary (i.e., the true upper limit) if any class is called the less than cumulative frequency of that class.

That table formed by writing the less than cumulative frequency of each class against the upper boundary (i.e. true upper limit) of the corresponding class is called the less than cumulative frequency distribution.

Now we give below the various steps involved in getting a less cumulative frequency distribution.

1. First of all we must see whether the given intervals are true class intervals or not. We must convert them into true class intervals.
2. Secondly, to the frequency of each class the frequencies of all the previous classes are to be added. That is, the frequency of the first class is left as it is, as there are no previous classes: to the frequency of the second class the frequency of the first class is added; to the frequency of the third class the frequencies of the first two classes are added and so on. The resultant figures are less than cumulative frequencies of the first, second, third, etc., classes respectively.
3. Now, in a table, the upper limits of the class intervals of the given simple frequency distribution are given under the heading 'less than'. The less than cumulative frequencies above are given under the heading 'frequency'. This table is the less than cumulative frequencies table.

Example - 1 :

Class interval	Frequency
0-15	2
15-30	5
30-50	13
50-60	8

From the above simple frequency distribution we get the less than cumulative frequency distribution as follows:

In the given frequency distribution the class intervals are true class intervals; so, there is no problem regarding conversion of given class intervals into true class intervals.

The frequency of the first class viz., 2 is kept as it is and it is the less than cumulative frequency of the first class.

To the frequency of the second class viz., 5 the frequency of the first class viz., 2 is added. The sum is $2 + 5 = 7$ and it is the less than cumulative frequency of the second class.

To the frequency of third class viz., 13 the frequencies of first two classes viz., 2 and 5 are add. The sum is $2 + 5 + 13 = 20$ and it is the less than cumulative frequency of the third class.

To the frequency of the fourth class viz., 8 the frequencies of the previous three classes viz., 2, 5 and 13 are added. The sum is $2 + 5 + 13 + 8 = 28$ and it is the less than cumulative frequency of the fourth class.

So the less than cumulative frequencies of the first, second, third and fourth classes are 2, 7, 20 and 28 respectively.

The true upper limits of the first, second, third and fourth classes are 15, 30, 50, 60 respectively.

Now, we give the less than cumulative frequency of each against the true upper limits of the corresponding class is follows:

Less than	Frequency
15	2
30	7
50	20
60	28

The above table is the required less than cumulative frequency table.

Example - 2 :

Class interval	Frequency
0-9	5
10-19	9
20-29	11
30-39	8
40-49	3

We get the less than cumulative frequency distribution from the above frequency distribution as follows:

In the given distribution the class intervals are not true class intervals. So,

we have to convert them into true class intervals. In the following table we have given the true class intervals.

Space for hints

True Class Interval	Frequency
- .5 - 9.5	5
9.5 - 19.5	9
19.5 - 29.5	11
29.5 - 39.5	8
39.5 - 49.5	3

We get the less than cumulative frequency as follows:

Less than cumulative frequency of the first class is 5, of the second class is $5 + 9 = 14$, of the third class is $5 + 9 + 11 = 25$. Of the fourth class is $5 + 9 + 11 + 8 = 33$ and of the fifth class is $5 + 9 + 11 + 8 + 3 = 36$.

The true upper limit of the first, second, third, fourth and fifth classes are 9.5, 19.5, 29.5, 39.5, 49.5 respectively.

Now, we give the less than cumulative frequency table as follows.

Class Interval	Frequency
9.5	5
19.5	14
29.5	25
39.5	33
49.5	36

It is to be noted that in the case of less than cumulative frequency distribution, less than cumulative frequency of the last class is equal to the total frequency.

(ii) Derivation of more than cumulative frequency distribution:

Here also, first of all, we give the definitions of 'more than cumulative frequency' and 'more than cumulative frequency distribution'.

When we are given a simple frequency distribution, number of items whose values are greater than the lower boundary (i.e. true lower limit) of any class is called the more than cumulative frequency of that class.

The table formed by writing the more than cumulative frequency of each class against the lower boundary (i.e. the true lower limit) of the corresponding class is called the more than cumulative frequency distribution.

Now, we give below the various steps involved in forming a more than cumulative frequency distribution.

1. First of all we must see whether the given class intervals are true class intervals or not. If not, we must convert them into true class intervals.
2. Secondly, to the frequency of each class the frequencies of all the succeeding classes are to be added. That is, to the frequency of the class the frequencies of all the remaining classes are added; to the frequency of the second class frequencies of the remaining classes, except the first class are added; to the frequency of the third class, frequencies of the remaining classes except the first two classes are added and so on. The frequency of the last class is left as it is as there is no succeeding class. The numbers obtained by the above method of addition of frequencies are more than cumulative frequencies of the first, second, third etc., classes respectively.
3. Now in a table, the true lower limits of the class intervals of the given frequency distribution are given under the heading 'more than'. The more than cumulative frequencies obtained above are given under the heading 'frequency'. This table is the more than cumulative frequency table.

Example - 1 :

Class interval	Frequency
0-15	2
15-30	5
30-50	13
50-70	8

From the above simple frequency distribution we get the more than cumulative frequency distribution as follows:

In the given frequency distribution the class intervals are true class intervals. So, there is no problem regarding conversion of given intervals into true class intervals.

The frequency of the first class is 2. To it the frequencies of the succeeding

classes (i.e.) the frequencies of the second, third and fourth classes viz., 5, 13 and 8 are added. The sum is $2 + 5 + 13 + 8 = 28$ and it is the more than cumulative frequency of the first class.

Space for hints

To the frequency of the second class viz., 5 the frequencies of the succeeding classes viz., 13 and 8 are added. The sum is $5 + 13 + 8 = 26$ and it is more than cumulative frequency of the second class.

To the frequency of the third class viz., 13 the frequency of the succeeding class viz, 8 is added. The sum is $13 + 8 = 21$ and it is more than cumulative frequency of the third class.

So, the more than cumulative frequencies of the first, second, third and fourth classes are 28, 26, 21 and 8 respectively.

The true lower limits of the first, second, third and fourth classes are 0, 15, 30, 50 respectively.

Now, we give the more than cumulative frequency of each class against the true lower limit of the corresponding class as follows.

More than	Frequency
0	28
15	26
30	21
50	8

The above table is the required more than cumulative frequency table.

Example - 2 :

Class Intervals	Frequency
0-9	5
10-19	9
20-29	11
30-39	8
40-49	3

We get the more than cumulative frequency distribution from the above frequency distribution as follows:

In the given distribution the class intervals are not true class intervals. So

we have to convert them into true class intervals. In the following table we have given the true class intervals.

Class Intervals	Frequency
-0.5 - 9.5	5
9.5 - 19.5	9
19.5 - 29.5	11
29.5 - 39.5	8
39.5 - 49.5	3

We get the more than cumulative frequencies as follows:

More than cumulative frequency of the first class is $5 + 9 + 11 + 8 + 3 = 36$, of the second class is $9 + 11 + 8 + 3 = 31$, of the third class is $11 + 8 + 3 = 22$, of the fourth class is $8 + 3 = 11$ and of the fifth class is 3.

The true lower limits of the first, second, third, fourth and fifth classes are -0.5, 9.5, 19.5, 29.5 and 39.5 respectively.

Now, we give the more than cumulative frequency table as follows:

More than	Frequency
-0.5	36
9.5	31
19.5	22
29.5	11
39.5	3

It is to be noted that in the case of more than cumulative frequency distribution more than cumulative frequency of the first class is equal to the total frequency.

5.11 Conversion of Cumulative frequency distribution into simple frequency distribution:

In case we are given less than cumulative frequency distribution or more than cumulative frequency distribution, they can be converted into simple frequency distribution. We have explained below the method of converting less than cumulative frequency distribution and more than cumulative distribution into simple frequency distribution.

(i) Methods of getting simple frequency distribution from less than cumulative frequency distribution:

Space for hints

In the case of less than cumulative frequency distribution, there are two columns-the first one bearing the heading 'less than' and second one containing less than cumulative frequencies. We have given below a less than cumulative frequency distribution.

Less than	Frequency
9.5	5
19.5	14
29.5	25
39.5	33
49.5	36

To convert this less than cumulative frequency distribution into simple frequency distribution, we have to (1) find out the different class intervals and (2) calculate simple frequencies from the given less than cumulative frequencies.

To find out the class intervals we require two of the following three pieces of information viz., (a) true upper limits (b) length of the class interval and (c) true lower limits. It is to be noted that the numbers given in the first column under the heading 'less than' are the true upper limit of the first class interval. Hence, in the less than cumulative frequency distribution given above, 9.5 is the true upper limit of the first class interval; 19.5 is the true upper limit of the second class interval; 29.5 is the true upper limit of the third class interval; 39.5 is the true upper limit of the fourth class interval; and 49.5 is the true upper limit of the fifth class interval.

Given the true upper limits of class intervals, we can find out the length of the class intervals. And given the true upper limits and length of the class intervals, we can find out the true lower limits of the class intervals. The method of finding out the true lower limits of the class intervals is explained below:

1. The difference between any two successive true upper limits gives us the length of the class interval. For instance, 9.5 and 19.5 are two successive true upper limits in the less than cumulative frequency distribution given above. Hence, the length of the class interval is 10.
2. Given the true upper limit and length of the class interval we can find out the true lower limit by subtracting the figure denoting the length of the class interval from the figure denoting the true upper limit. For instance, in the less than

cumulative frequency distribution given above, we come to know that the true upper limit of the first class interval is 9.5. And we know that the length of the class interval is 10. Hence, by subtracting 10 from 9.5 we can get the true lower limit. That is, the true lower limit of the first class interval is equal to 9.5 minus 10 viz. -0.5 . Hence, the first class interval is $-0.5 - 9.5$.

In this way, we can find out other class intervals also

$$\begin{aligned}\text{Lower limit of the second class interval} &= \text{Upper limit of the second class interval} - \text{Length of the class interval} \\ &= 19.5 - 10 = 9.5\end{aligned}$$

Hence, the second class interval is $9.5 - 19.5$

$$\begin{aligned}\text{Lower limit of the third class interval} &= \text{Upper limit of the third class interval} - \text{Length of the class interval} \\ &= 39.5 - 10 = 29.5\end{aligned}$$

Hence, the fourth class interval is $29.5 - 39.5$

$$\begin{aligned}\text{Lower limit of the fifth class interval} &= \text{Upper limit of the fifth class interval} - \text{Length of the class interval} \\ &= 49.5 - 10 = 39.5\end{aligned}$$

Hence, the fifth class interval is $39.5 - 49.5$.

Thus, we get the following class intervals:

Class Interval
$-0.5 - 9.5$
$9.5 - 19.5$
$19.5 - 29.5$
$29.5 - 39.5$
$39.5 - 49.5$

The next task is the calculation of simple frequencies from the less than cumulative frequencies given.

When we are given a less than cumulative frequency distribution we need not adopt any method to find out the simple frequency of the first class. So far as the first class is concerned, the less than cumulative frequency distribution given above, the less than cumulative frequency of the first class is 5 and it is the simple frequency of the first class.

The simple frequencies of the other classes are calculated by adopting the following method. So far as the other classes are concerned simple frequency of a class is equal to the difference between the less than cumulative frequency of

that class and the less than cumulative frequency of the immediately preceding class. For instance, in the less than frequency distribution given above, less than cumulative frequency of the second class is 14 and that for the first class is 5. The difference between these two is 9. Hence, the simple frequency of the second class interval (viz., 9.5 -19.5) is 9.

In our example, the less than cumulative frequency of the third class is 25 and that of the second class is 14. The difference between these two is 11. Hence, the simple frequency of the third class interval (viz., 19.5-29.5) is 11.

The less than cumulative frequency of the fourth class is 33 and that of the third class is 25. The difference between these two is 8. Hence, the simple frequency of the fourth class interval (viz., 29.5-39.5) is 8.

The less than cumulative frequency of the fifth class is 36 and that of the fourth class is 33. The difference between these two is 3. Hence, the simple frequency of the fifth class interval (viz., 39.5-49.5) is 3. Thus we get the following simple frequency table for the less than cumulative frequency distribution given above is as follows:

True Class interval	Frequency
-0.5 - 9.5	5
9.5 - 19.5	9
19.5 - 29.5	11
29.5 - 39.5	8
39.5 - 49.5	3

(ii) Method of getting simple frequency distribution from more than cumulative frequency distribution:

In the case of more than cumulative frequency distribution there are two columns the first one bearing the heading 'more than' and the second one containing more than cumulative frequencies. We have given below a more than cumulative frequency distribution.

More than	Frequency
9.5	5
19.5	14
29.5	25
39.5	33
49.5	56

To convert this more than cumulative frequency distribution into simple frequency distribution, we have to (1) find out the different class intervals (2) calculate simple frequencies from the given more than cumulative frequencies.

To find out the class intervals we require two of the following three pieces of information viz., (a) true lower limits (b) length of the class interval and (c) true upper limits. It is to be noted that the numbers given in the first column under the heading 'more than' are the true lower limits of class intervals. Hence, in the more than cumulative frequency distribution given above, -0.5 is the true lower limit of the first class interval; 9.5 is the true lower limit of the second class interval; 19.5 is the true lower limit of the third class interval; 29.5 is the true lower limit of the fourth class interval; and 39.5 is the true lower limit of the fifth class interval.

Given the true lower limits of class intervals, we can find out the length of class intervals. And given the true lower limits and length of the class intervals, we can find out the true upper limit of the class intervals. The method of finding out the upper limits of class intervals is explained below :

1. The difference between any two successive true lower limits gives us the length of the class interval. For instance, -0.5 and 9.5 are two successive true lower limits in the more than cumulative frequency distribution given above. Hence the length of the class intervals is 10 .

2. Given the true lower limit and length of the class interval, we can find out the true upper limit by adding the figure denoting the length of the class interval with the figure denoting the true lower limit. For instance, in the more than cumulative frequency distribution given above, we know that the true lower limit of the first class interval is -0.5 . And we know that the length of the class interval is 10 . Hence, by adding 10 with -0.5 , we can get the true upper limit. That is, true upper limit of the first class interval is equal to -0.5 plus 10 viz., 9.5 . Hence, the first class interval is $-0.5 - 9.5$.

In this way we can find out the other class intervals also.

Upper limit of the second class interval = Lower limit of the second class interval
+ Length of the class interval.

$$= 9.5 + 10 = 19.5$$

Hence, the second class interval is $9.5 - 19.5$

Upper limit of the third class interval = Lower limit of the third class interval +
Length of the class interval.

$$= 19.5 + 10 = 29.5$$

Hence, the third class interval is 19.5 - 29.5

Upper limit of the fourth class interval = Lower limit of the fourth class interval
+ Length of the class interval.

$$= 29.5 + 10 = 39.5$$

Hence, the fourth class interval is 29.5 - 39.5

Upper limit of the fifth class interval = Lower limit of the fifth class interval
+ Length of the class interval.

$$= 39.5 + 10 = 49.5$$

Hence, the fifth class interval is 39.5 - 49.5.

Thus, we get the following class intervals:

Class Intervals		
-0.5	-	9.5
9.5	-	19.5
19.5	-	29.5
29.5	-	39.5
39.5	-	49.5

The next task is the calculation of simple frequencies from the more than cumulative frequencies given.

When we are given a more than cumulative frequency distribution, we need not adopt any method to find out the simple frequency of the last class. So far as the last class is concerned, the more than cumulative frequency is the simple frequency of that class. For instance, in the more than cumulative frequency distribution given above, the more than cumulative frequency of the last class is 3. Hence, this is simple frequency of the last class viz., 39.5 - 49.5) is 3.

The simple frequencies of the other classes are calculated by adopting the following method. So far as the other classes are concerned simply frequency of a class is equal to the difference between the more than cumulative frequency of that class and of the class immediately succeeding it. For instance, in the more than cumulative frequency distribution given above, more than cumulative frequency of the first class is 36 and that of the second class is 31. The difference

between these is 5. Hence the simple frequency of the first class interval (viz. $-5 - 9.5$) is 5.

In our example, The more than cumulative frequency of the second class is 31 and that of third class interval is 22. The difference between these two is 9. Hence, the simple frequency of the second class interval (viz. $9.5 - 19.5$) is 9.

The more than cumulative frequency of the third class is 22 and that of the fourth class is 11. The difference between these two is 11. Hence, the simple frequency of the third class interval (viz. $19.5 - 29.5$) is 11.

The more than cumulative frequency of the fourth class is 11 and that of the fifth class is 3. The difference between these two is 8. Hence the simple frequency of the fourth class interval (viz., $29.5 - 39.5$) is 8.

[As stated above the simple frequency of the last class is the same as the more than cumulative frequency of that class. In our illustration, the fifth class is the last class (viz, $39.5 - 49.5$) and its frequency is 3..

Thus we get the following simple frequency table for the given more than cumulative frequency table given below :

Class Interval	Frequency
$-0.5 - 9.5$	5
$9.5 - 19.5$	9
$19.5 - 29.5$	11
$29.5 - 39.5$	8
$39.5 - 49.5$	3

Example:

Form a frequency distribution with appropriate class intervals from the following data:

56, 24, 89, 42, 56, 72, 91, 96, 43, 32, 19, 62, 75, 66, 54, 48, 52, 82, 36, 62, 41, 37, 85, 72, 66, 54, 34, 51, 27, 39, 68, 53, 74, 81, 29, 61, 49, 36, 86, 81.

Derive the less than and more than cumulative frequency distribution from it.

Answer

In the given set of data maximum value is 96 and minimum value is 12.

Let us choose the class intervals to be 10-29, 30-49, 50-69, 70-89 and 90 - 9.

Now we give the frequency table as follows:

Space for hints

Class Intervals	Tally marks	Frequency
10-29		4
30-49		11
50-69		13
70-89		10
90-109		2

We have to derive the less than and more than cumulative frequency distribution from the above frequency table; for that, first of all we convert the above class intervals into true class intervals.

True class Intervals	Frequency
9.5-29.5	4
29.5-49.5	11
49.5-69.5	13
69.5-89.5	10
89.5-109.5	2

The less than cumulative frequency distribution is as follows:

Less than	Frequency
29.5	4
49.5	15
69.5	18
89.5	38
109.5	40

The more than cumulative frequency distribution is as follows:

More than	Frequency
9.5	40
29.5	36
49.5	25
69.5	12
89.5	2

Example:

A continuous frequency, distribution is given below:

Class interval	0-5	6-15	16-30	31-40	41-45
Frequency	3	8	10	5	4

- What is the length of each class interval?
- What are the midvalues of the class intervals?
- Derive the less than and more than cumulative frequency distribution.

The given class intervals are not true class intervals. To find out the length of each class first we convert them into true class intervals.

True Class	-0.5-5.5	5.5-15.5	15.5-30.5	30.5-40.5	40.5-45.5
Frequency	3	8	10	5	4

Now, the lengths of first, second, etc., classes are 6, 10, 15, 10 and 5 respectively.

The midvalues of the first, second etc., classes are 2.5, 10.5, 23, 35.5 and 43 respectively.

The less than cumulative frequency distribution is as follows:

Less than	Frequency
5.5	3
15.5	11
30.5	21
40.5	26
45.5	30

The more than cumulative frequency distribution is as follows:

More than	Frequency
-0.5	30
5.5	27
15.5	19
30.5	9
40.5	4

Example:

Space for hints

Derive the less than and more than cumulative frequency distribution from the following frequency table:

Midvalue	5	15	25	35	45	55
Frequency	4	7	11	12	8	3

To derive the cumulative frequency distribution, first of all, we find out the class intervals from the midvalues:

Class interval	Frequency
0 -10	4
10-20	7
20-30	11
30-40	12
40-50	8
50-60	3

The less than and more than cumulative frequency distributions are as follows:

Less than	Frequency
10	4
20	11
30	22
40	34
50	42
60	45

More than	Frequency
0	45
10	41
20	34
30	23
40	11
50	3

6. Answers to the Check Your Progress Questions :

- | | |
|-------------------|---------------|
| 1. Refer 1.1 | 8. Refer 3.1 |
| 2. Refer 1.2 | 9. Refer 4.1 |
| 3. Refer 1.4 | 10. Refer 4.2 |
| 4. Refer 1.10 | 11. Refer 4.9 |
| 5. Refer 2.3(iii) | 12. Refer 4.9 |
| 6. Refer 2.3(iv) | 13. Refer 5.1 |
| 7. Refer 3.1 | 14. Refer 5.7 |

7. Model questions for guidance :**10 Marks Questions (One Page Answer)**

1. Explain the use of statistical analysis for studies in economics.
2. "A knowledge of statistics is like a knowledge of foreign language or of algebra; it may prove of use at any time under any circumstances" - Comment.
3. Explain the application of statistics in business.
4. Account for the distrust in Statistics
5. What are the characteristics of Statistics?
6. Write a short note on limitations of Statistics.
7. Examine the validity of the following statement: Statistics are white lies.
8. What do you mean by secondary data? State their important sources and explain what precautions are necessary before using them.
9. What are the merits and demerits of secondary data?
10. What are the merits and defects of collecting primary data?
11. Distinguish between primary and secondary data.
12. Distinguish between Questionnaire and schedule.
13. Write short notes on Questionnaire.

20 Marks Questions (Three Page Answer)

1. Statistics is the science of measurement of the social organism regarded as a whole in all its manifestations", (Bowley). Discuss this definition with

2. Statistics is a 'science of counting', Discuss.
3. Statistics is a 'science of average', Discuss.
4. Give an account of the important definitions of statistics, pointing out the one you think the best.
5. "Statistics is a science which deals with estimates rather than with exact enumerations". Examine the statement.
6. What are the characteristics that statistics (statistical data) possess. Explain with illustrations.
7. Explain the characteristics and limitations of statistics.
8. Explain the usefulness of statistics in the different fields of enquiry.
9. What do you understand by Statistical methods? Discuss the scope, utility and limitations of these methods. Give an example of misuse of statistics.
10. Explain the uses and limitations of statistical methods.
11. Explain the relationship of statistics with other sciences with particular reference to Economics.
12. Define Statistics and discuss its limitations.
13. What are the various methods of collecting primary data? Briefly state their merits and demerits.
14. What do you mean by a questionnaire? State the essentials of a good questionnaire and explain the merits and defects of using it for collecting statistical data.
15. Compare the different methods used in the collection of Statistical data.

UNIT - 2

AVERAGES

Introduction

Whether literate or illiterate, everybody is familiar with the term 'average'. Average is a fundamental statistical measure. There are different types of average and some averages are more popular and widely used by all. We explain the meaning and method of computation of five popular averages namely, Arithmetic Mean, Median, Mode, Geometric Mean and Harmonic Mean in this Unit-2. In the last section of this unit, we compare the different averages in respect of their merits and demerits

Unit Objectives :

After studying this Unit, you would be able to understand

- * the meaning and computational technique with reference to mean, median, mode, geometric mean and harmonic mean.
- * the relative merits and demerits of various averages.

Unit Structure :

1. Averages - Meaning, Objectives and Types
2. Arithmetic Mean
3. Median
4. Quartiles
5. Mode
6. Geometric Mean
7. Harmonic Mean
8. Relative Merits and Demerits of various Averages
9. Answers to the Check Your Progress Questions
10. Model questions for guidance

1. AVERAGES - Meaning, Objectives and Types

1.1 Meaning

We have discussed in the earlier topics how large mass of data are condensed into frequency tables. Various distributions given in the form of tables cannot be compared directly. Suppose, two tables of figures of marks obtained by 100 students belonging to two Universities are given; it would

be impossible to arrive at any conclusion, if the figures given in two tables are compared directly. Hence, in order to make comparisons and to draw conclusions from a given set of data, it is necessary to have some single measurement which may describe the characteristic of the given data. Often it happens that in the given distributions some values occur more frequently, and other values occur less frequently. The value which occurs most frequently usually lies in the central part of the distribution. The values calculated to measure this characteristic of the distributions are called '**measures of central tendency**' or '**averages**'.

Measures of central tendency are also called '**types**' as they are typical values of a series. Since averages locate a distribution at some value of the variable, they are sometimes known as '**measures of location**'.

An average is expressed necessarily in the same unit in which the series is. For example, if the value of a given variable is expressed in rupees, average will also be given in rupees.

1.2 Objectives :

The main object of an average is to present huge mass of statistical data in a simple and concise manner. This makes the central theme of the data readily intelligible.

1.3 Types of averages :

There are several types of averages. The following are the main types of averages that are commonly used.

1. Arithmetic mean

2. Median

3. Mode

4. Geometric mean

5. Harmonic mean

1.4 Characteristics of a good or representative statistical average

Each average has its own merits and demerits and is used for different purposes. No single average is suitable for all purposes; hence, a selection has to be made regarding the average most suitable for the purpose at hand. The average which satisfies most of the following characteristics can be considered to be the best for the purpose at hand.

- 1) It should be calculated by a rigidly defined formula and not be a product of guesswork or estimation.
- 2) It should properly be based on the values of all the items in the distribution.
- 3) Its value should not be usually affected by the influence of extremely high or low values. If, irrespective of the values of other items, the presence of some extremely high values make value of the average to be large and the presence of some extremely low values make the average value small, then the average cannot be considered to be typical.
- 4) The calculations involved in finding it, should be simple and easy to understand. An average is expected to simplify complexities. Hence, its meaning should not be complex and ambiguous. It should be calculated with ease and rapidity.
- 5) It should lend itself for further algebraic treatment. That is, it should be flexible in other uses. Suppose the separate averages of heights of male and female students of a particular class are given. It should be possible to find out the combined average of heights of both male and female students of the class with the help of the separate averages.
- 6) It should be stable. The value of the average should not be affected much by small changes in the method of classification or by small errors of observations. Suppose, from a given population two or more samples are drawn. We can calculate the value of the average from each sample. The values obtained may not all be equal. There may be differences and these differences are called fluctuations of sampling. Under such conditions, averages having no appreciable differences are considered to be stable and representative.

2. ARITHMETIC MEAN

2.1 Definition

Arithmetic mean is generally known as 'Mean' or the common 'average'. It is defined as the total value of all the items given divided by the total number of items given.

Check your Progress

1. What is Arithmetic mean?

Mean can be calculated for (i) ungrouped data (ii) discrete frequency distributions (iii) continuous frequency distributions.

Space for hints

2.2 Mean for ungrouped data :

(i) Direct Method :

The arithmetic mean of a given set of items is obtained by summing up the values of all items and dividing the sum by the total number of items.

Let us suppose that the given set of items is the age of 5 students viz. 17, 16, 15, 18 and 16. We use the symbol x_1 to denote the age 17, x_2 to denote the age 16, x_3 to denote the age 15, x_4 to denote the age 18 and x_5 to denote the age 16.

To denote the total number of items in any given set the letter 'n' is used. In the illustration given above, the total number of items is 5 and therefore $n = 5$.

To denote the total value of all items in the given set (e.g. : $17+16+15+18+16$) we use the symbol Σx . In this, Σ represents 'summation of' and is read as sigma.

Mean is usually denoted by the symbol \bar{x} (which is read as x bar).

$$\text{Mean } (\bar{x}) = \frac{\text{Total value of all the items given } (\Sigma x)}{\text{The total number of items given } (n)}$$

$$\therefore \bar{x} = \frac{\Sigma x}{n}$$

In the illustration we have given above

$$\Sigma x = 17+16+15+18+16 = 82$$

$$n = 5$$

$$\therefore \bar{x} = \frac{82}{5} = 16.4$$

Calculating mean using the formula $\bar{x} = \frac{\Sigma x}{n}$ is called direct method.

Check your Progress

2. What is the formula to calculate \bar{x} for ungrouped data?

Example 1 :

A set of items are given below :

30, 80, 20, 75, 300, 10, 8, 42, 250, 36, 40. Calculate the mean.

Total value of the items given = Σx

$$= 30 + 80 + 20 + 75 + 300 + 10 + 8 + 42 + 250 + 36 + 40 = 891$$

Total number of items = $n = 11$.

$$\begin{aligned} \therefore \bar{x} &= \frac{\Sigma x}{n} \\ &= \frac{891}{11} = 81 \end{aligned}$$

Example 2 :

Marks obtained by students in an examination are as follows :

Roll No.	1	2	3	4	5	6	7	8	9	10
Marks	93	75	80	45	60	30	42	45	50	46

Calculate the mean.

n = Total number of items = 10

Total value of the items given = Σx

$$= 93 + 75 + 80 + 45 + 60 + 30 + 42 + 45 + 50 + 46 = 566$$

$$\begin{aligned} \therefore \bar{x} &= \frac{\Sigma x}{n} \\ &= \frac{566}{10} = 56.6 \end{aligned}$$

Answer : Mean = 56.6

(ii) Short-cut Method :

If we have a moderate number of items and small sized figures, then the mean may be calculated by the method given above (known as direct method). If the number of items is large and the values of the items are large,

then, the process of adding together all the values may be a lengthy one. To remove this lengthy process short-cut method is used and this method is explained below.

Space for hints

Consider the same example giving the ages of 5 students viz., 17, 16, 15, 18 and 16.

Some value within the range of given set of items is chosen as origin and is denoted by A. In the example given above the range is between 15 and 18. We can take any value between 15 and 18 as origin (i.e., A). Let A = 16.

Now the deviation of each item from A is calculated. It is denoted by d. If x is the value of one item then $d = (x - A)$. In the illustration given above, the values of d are (17-16), (16-16), (15-16), (18-16) and (16-16), i.e., 1, 0, -1, 2 and 0.

Total value of the deviations (e.g., 1+0-1+2+0) is denoted by Σd .

Now, the mean is given by the following formula.

$$\bar{x} = A + \frac{\Sigma d}{n}$$

Where \bar{x} = Mean

A = Arbitrary origin chosen

Σd = total sum of deviations of all items from A

n = Total number of items

In the illustration given above A = 16.

$$\Sigma d = 1+0-1+2+0 = 2 \text{ and } n = 5$$

$$\therefore \bar{x} = 16 + 2/5 = 16.4$$

Answer : **Mean = 16.4**

Example 2 :

We can solve the problem given under Example-I by the short-cut method and the method of calculation is given below :

We can choose the origin A to be equal to 70.

The set of items is 30, 80, 20, 75, 300, 10, 8, 42, 250, 36, 40.

The deviation of the first item from A is $30-70 = -40$ deviation of the second item from A is $80-70 = 10$ and so on.

The deviation of the last item from A is $40 - 70 = -30$

$\therefore \Sigma d =$ Total sum of deviations of all the items from A.

$$= -40+10-50+5+230-60-62-28+180-34-30$$

$$= -40-50-60-62-28-34-30+10+5+230+180$$

$$= -304+425 = 121$$

$$n = \text{Total number of items} = 11$$

$$\therefore \text{Mean} = \bar{x} = A + \frac{\Sigma d}{n}$$

$$= 70 + \frac{121}{11} = 70+11 = 81$$

Answer : **Mean = 81**

Note that the value of the mean got by the short cut method is the same as the value got by the direct method.

Example 3 :

Twelve values of a variable x are given below. Calculate the mean.

x : 85, 79, 75, 78, 90, 86, 74, 65, 73, 54, 40 and 35.

We can choose the origin A to be equal to 60

(x-A) : (85-60), (79-60), (75-60), (78-60), (90-60), (86-60), (74-60), (65-60), (73-60), (54-60), (40-60) and (35-60).

Hence the values of d are given as follows :

d : 25, 19, 15, 18, 30, 26, 14, 5, 13, -6, -20 and -25

$\Sigma d =$ Total value of deviations of all the items from A

$$= 25+19+15+18+30+26+14+5+13-6-20-25$$

$$= 165 - 51 = 114$$

$$n = \text{Total number of items} = 12$$

$$\begin{aligned} \therefore \text{Mean} = \bar{x} &= A + \frac{\sum d}{n} \\ &= 60 + \frac{114}{12} = 60 + 9.5 = 69.5 \end{aligned}$$

Answer : **Mean = 69.5**

2.3 Mean of a Discrete Frequency Distribution

(i) Direct Method :

Consider the following discrete frequency distribution.

Marks (x)	Frequency (f)
45	3
53	6
54	2
55	1

We use the symbol x_1 to denote the mark 45, x_2 to denote the mark 53, x_3 to denote the mark 54, x_4 to denote the mark 55.

Frequency 3 is denoted by f_1 , 6 by f_2 , 2 by f_3 , and 1 by f_4

f_1 items are having the value x_1 and hence the total value of the f_1 items = $x_1 f_1$

f_2 items are having the value x_2 and hence the total value of the f_2 items = $x_2 f_2$

f_3 items are having the value x_3 and hence the total value of the f_3 items = $x_3 f_3$

f_4 items are having the value x_4 and hence the total value of the f_4 items = $x_4 f_4$

\therefore Total value of all the items given

$$= x_1 f_1 + x_2 f_2 + x_3 f_3 + x_4 f_4$$

Check your Progress

3. What is the formula to calculate mean for Discrete Frequency Distribution?

This total is denoted by the symbol Σxf . From the example given above

$$\begin{aligned}\Sigma xf &= (45 \times 3) + (53 \times 6) + (54 \times 2) + (55 \times 1) \\ &= 135 + 318 + 108 + 55 = 616\end{aligned}$$

Total number of items viz., $f_1 + f_2 + f_3 + f_4$ is denoted by the symbol Σf . In the example given above.

$$\Sigma f = 3 + 6 + 2 + 1 = 12$$

Mean is defined as the total value of all items divided by the total number of items. Hence, we can give the formula as follows :

$$\bar{x} = \frac{\Sigma xf}{\Sigma f}$$

where \bar{x} = Mean

Σxf = Total value of all the items

Σf = Total number of items

In the example given above, $\Sigma xf = 616$ and $\Sigma f = 12$

$$\therefore \bar{x} = \frac{616}{12}$$

$$= 51.33 \text{ (approx.)}$$

Answer : **Mean = 51.33**

In the case of discrete frequency distribution, the method of calculation of mean as described above is known as 'direct method'.

Example 4 :

Calculate the mean of the following discrete frequency distribution.

Size of item (x)	1	2	3	4	5	6	7	8	9	10	11	12
Frequency (f)	2	9	25	50	60	71	45	28	22	15	9	1

The required computations for mean calculations are done in the table below.

Space for hints

x	f	xf
1	2	2
2	9	18
3	25	75
4	50	200
5	60	300
6	71	426
7	45	315
8	28	224
9	22	198
10	15	150
11	9	99
12	1	12
Total	$\Sigma f = 337$	$\Sigma xf = 2019$

$$\bar{x} = \frac{\Sigma xf}{\Sigma f}$$

$$= \frac{2019}{337} = 5.99 \text{ (approximately)}$$

Answer

Mean = 5.99

Example 5 :

From the following data find out the value of the mean.

Income (Rs.) x	18	80	100	150	200	250
No. of persons f	30	16	24	26	20	6

x	f	xf
18	30	540
80	16	1280
100	24	2400
150	26	3900
200	20	4000
250	6	1500
Total	$\Sigma f = 122$	$\Sigma xf = 13620$

$$\begin{aligned}\bar{x} &= \frac{\Sigma xf}{\Sigma f} \\ &= \text{Rs. } \frac{13620}{122} = \text{Rs. } 111.64 \text{ (approx.)}\end{aligned}$$

Answer :

Mean = Rs.111.64

(ii) Short-cut Method

As we have used short-cut method in the calculation of mean for ungrouped data, here in the case of discrete frequency distribution also we have a short-cut method. This method consists of the following steps :

1. Any arbitrary value usually, the value of the item corresponding to the maximum frequency is chosen as origin. It is denoted by A. The arbitrary value (viz., A) is also called the working mean.
2. Deviation of each value given from A is calculated. If x is one of the given values then its deviation from A is (x-A). This deviation is denoted by d.
3. These deviations are multiplied by the corresponding frequencies. If f is the frequency corresponding to the deviation d then f and d are multiplied. These products are summed up and the sum is denoted by Σfd .
4. To get the mean value, the above sum Σfd is divided by the total number of items Σf and is added to A.

$$\therefore \bar{x} = A + \frac{\sum fd}{\sum f}$$

Consider the discrete frequency distribution given earlier viz.

Marks (x)	Frequency (f)
45	3
53	6
54	2
55	1

Let us choose the origin A to be equal to 50.

Then the deviation of each value from the arbitrary value A is calculated. In the example given above, the deviation of the first value viz., 45 from A is $45-50=-5$. The deviation of the second value from A is $53-50=3$. Similarly, the deviations of third and fourth values are $4(54-50)$ and $5(55-50)$ respectively. Hence, the values of d are as follows; d : -5, 3, 4 and 5.

Now, these deviations are multiplied by the corresponding frequencies. In our example, deviation (-5) is multiplied by the frequency 3, deviation 3 by the frequency 6, deviation 4 by the frequency 2 and deviation 5 by the frequency 1. Therefore the values of fd are as follows

$$fd : [(-5) \times 3], (3 \times 6), (4 \times 2) \text{ and } (5 \times 1)$$

The products we have got above are summed up and the sum gives the value of $\sum fd$.

$$\begin{aligned} \sum fd &= [(-5) \times 3] + (3 \times 6) + (4 \times 2) + (5 \times 1) \\ &= -15 + 18 + 8 + 5 \\ &= -15 + 31 = 16 \end{aligned}$$

As we have noted earlier, total number of items = $f_1 + f_2 + f_3 + f_4 = \sum f$.

Thus the mean value is given by the following formula.

$$\bar{x} = A + \frac{\sum fd}{\sum f}$$

where \bar{x} = Mean

A = Arbitrary origin

Σfd = Total sum of deviations of all the items from A

Σf = Total number of items

In our example, $A = 50$, $\Sigma fd = 16$, $\Sigma f = 12$

$$\therefore \bar{x} = 50 + \frac{16}{12} = 50 + 1.33 \text{ (approx.)} = 51.33$$

Answer : **Mean = 51.33**

Note that the mean value obtained by the short-cut method and by the direct method are the same.

Example 6 :

Consider the problem given under Example 5. We can find out the mean by the short-cut method as follows : Let us take A to be equal to 150.

$$\bar{x} = A + \frac{\Sigma fd}{\Sigma f}$$

The required computations are made in the Table below:

x	f	$d = (x - A)$ $= (x - 150)$	fd
18	30	-132	-3960
80	16	-70	-1120
100	24	-50	-1200
150	26	0	0
200	20	50	1000
250	6	100	600
Total	$\Sigma f = 122$		$\Sigma fd = -4680$

$$[\Sigma fd = -3960 - 1120 - 1200 + 0 + 1000 + 600$$

$$= -6280 + 1600$$

$$= -4680]$$

$$\bar{x} = 150 + \frac{(-4680)}{122}$$

$$= 150 - \frac{4680}{122} = 150 - 38.36 = 111.64$$

Space for hints

Answer : **Mean = 111.64**

Example 7 :

Consider the table given below. Calculate the mean by the short-cut method.

Wages in Rs.(x)	4	6	8	10	15	17	18	20	21	22
No. of workers (f)	5	20	16	10	6	7	8	5	3	2

Let us take 7 to be the origin A.

$$\bar{x} = A + \frac{\sum fd}{\sum f}$$

x	f	d = (x-A) = (x-7)	fd
4	5	-3	-15
6	20	-1	-20
8	16	1	16
10	10	3	30
15	6	8	48
17	7	10	70
18	8	11	88
20	5	13	65
21	3	14	42
22	2	15	30
Total	$\Sigma f = 82$		$\Sigma fd = 354$

$$\bar{x} = A + \frac{\sum fd}{\sum f}$$

$$= 7 + \frac{354}{82}$$

$$= 7 + 4.32 \text{ (approx.)}$$

$$= 11.32$$

Answer : **Mean = 11.32**

2.4 Mean of continuous frequency distribution :

(i) Direct Method

Consider the following continuous frequency distribution :

Class Interval(Marks)	Frequency (No. of Students)
40 - 44	9
45 - 49	11
50 - 54	10
55 - 59	5

In a continuous frequency distribution the exact values of the items are not given; only the range within which each item has its value is indicated. In the distribution given above 9 students are having their marks within the range 40 to 44; 11 students have their marks within the range 45 to 49; marks of 10 students fall within the range 50 to 54; and 5 students have their marks within the range 55 to 59. The exact value of marks obtained by each student is not known. Hence, it is not possible to find out the value of the marks obtained by all the students.

But the arithmetic mean is always defined as the total value of all the items divided by the total number of items. Hence, it is necessary to calculate the total value of all the items given. In order to calculate the total value of all the items in a continuous frequency distribution we adopt the following rule:

Each class interval of the continuous frequency distribution is replaced by its mid value. By this we mean that all the items in an interval are having their values equal to the mid value of that class.

In our example, the interval 40-44 is replaced by its mid-value 42; interval 45-49 is replaced by its mid-value 47; 50-54 is replaced by its mid-value 52. The mid-value of any class represents the marks obtained by the students of that class. Therefore by the above replacements we mean that the marks obtained by each of the 9 students belonging to the first class = 42;

marks obtained by each of the 11 students in the second class = 47; marks obtained by each of the 10 students in the 3rd class = 52; and the marks obtained by each of the 5 students in the last class = 57.

Space for hints

The mid-value of the first class (e.g. 42) is denoted by x_1 . Mid value of the second class (e.g. 47) is denoted by x_2 ; and so on.

The frequency of the first class (e.g. 9) is denoted by f_1 ; frequency of the second class (e.g. 11) is denoted by f_2 ; and so on.

The total value of all the items in the first class is obtained by multiplying x_1 by f_1 . Therefore, $x_1 f_1$ gives the total value of all the items in the first class. Similarly $x_2 f_2$ gives the total value of all the items in the second class and so on.

In our example, 42×9 is the total value of all the items in the first class. 47×11 is the total value of all the items in the second class. 52×10 is the total value of all the items in the third class and 57×5 is the total value of all the items in the fourth class.

Total value of all the items in the distribution as a whole is got by summing the total values of items in all the classes, and hence is equal to $x_1 f_1 + x_2 f_2 + \dots$

This is denoted by $\sum xf$

$$\begin{aligned} \text{In our example } \sum xf &= (42 \times 9) + (47 \times 11) + (52 \times 10) + (57 \times 5) \\ &= 378 + 517 + 520 + 285 = 1700 \end{aligned}$$

Total number of items viz., $f_1 + f_2 + \dots$ is denoted by $\sum f$

$$\text{In our example } \sum f = 9 + 11 + 10 + 5 = 35$$

Now, the mean is given as follows :

$$\text{Mean } \bar{x} = \frac{\text{Total value of all the items } (\sum xf)}{\text{Total number of items } (\sum f)}$$

$$\bar{x} = \frac{\sum xf}{\sum f}$$

Where $\sum xf$ = sum of the products got by multiplying the midvalue of each class by the corresponding frequency.

$$\sum f = \text{Total number of items}$$

For the example given above,

$$\bar{x} = \frac{1700}{35} = \frac{340}{7} = 48.57$$

Answer : **Mean = 48.57**

The method described above is called direct method.

Example 8 :

Calculate the mean of the following data :

Class	Frequency
0 - 6	4
6 - 12	8
12 - 18	14
18 - 24	16
24 - 30	20

The mid value of each class is found out. The mid-value of the first class is 3; mid value of the 2nd class is 9 and so on. The class intervals are replaced by these mid values.

Then each mid value is multiplied by the corresponding frequency. Mid value 3 is multiplied by the frequency 4, mid value 9 is multiplied by the frequency 8, and so on.

The sum of these products is calculated and this gives the value of $\sum xf$.

The sum of all the frequencies gives $\sum f$.

\therefore Mean $\bar{x} = \frac{\sum xf}{\sum f}$ can be calculated.

Mid value x	Frequency f	xf
3	4	12
9	8	72
15	14	210
21	16	336
27	20	540
Total	62	1170

$$\bar{x} = \frac{1170}{62} = 18.87 \text{ (approximately)}$$

Answer : **Mean = 18.87**

(ii) Short-cut Method :

Space for hints

A short-cut method of calculating mean in the case of continuous frequency distribution is explained below. There are two types of continuous frequency distribution viz., (A) continuous frequency distribution with equal class intervals and (B) continuous frequency distribution with unequal class intervals. First the method of calculation of mean in the case of continuous frequency distribution with equal class intervals is given below.

(A) Short-cut method in the case of continuous frequency distribution with EQUAL CLASS INTERVALS

(1) First the mid values of all the class intervals are found out. Any one of the mid values is chosen as origin A which is also known as the working mean. Usually the working mean A is chosen from the central class in the distribution. If there are two central classes in the distribution then the mid value of the central class having greater frequency is chosen as origin. Consider the following distribution.

Distribution I :

Class interval	Mid value x	Frequency f
0 - 10	5	2
10 - 20	15	18
20 - 30	25	45
30 - 40	35	20
40 - 50	45	6

There are 5 classes in the above distribution. Hence, the mid value of the third class will be chosen as origin. Thus, A will be equal to 25.

Instead of the above distribution let us suppose that the following is the given distribution.

Distribution II :

Class interval	Mid value x	Frequency f
0 - 10	5	2
10 - 20	15	18
20 - 30	25	40
30 - 40	35	45
40 - 50	45	20
50 - 60	55	6

Check your Progress

4. What is the formula to calculate Mean for continuous frequency distribution?

In the above distribution there are 6 classes and hence we have two central classes viz., the 3rd and 4th class. But the frequency of the 4th class (viz., 45) is greater than the frequency of the third class (viz., 40). The midvalue of the central class corresponding to the greater frequency is 35, and this midvalue (viz., 35) is chosen as origin A.

(2) Once the origin A is chosen, the deviation of the midvalue of each class from A is calculated. If X is one of the midvalues then $(x-A)$ is its deviation from A.

For the distribution I, $A = 25$. The deviation of the midvalue 5 from A is $5-25 = -20$; deviation of the midvalue 15 from A is $15-25 = -10$; and so on.

The deviations are $(-20), (-10), 0, 10$, and 20 .

(3) The deviation of each midvalue from A is divided by the 'length of the class interval'. If 'c' denotes the length of the class interval, then each deviation $(x-A)$ is divided by c.

We denote $\left(\frac{x-A}{c}\right)$ by d.

Consider distribution I given above having 5 classes. All the five classes are uniform and they have the same width or length which is equal to 10.

$$\therefore c = 10$$

Hence, deviation of each midvalue from the origin A is divided by 10 and the values of d are obtained.

In our example, the deviations of midvalues from A are $-20, -10, 0, 10$ and 20 . These deviations are divided by 10.

Hence, the values of d are

$$-2 \left[\text{i.e., } \frac{-20}{10} \right], -1 \left[\text{i.e., } \frac{-10}{10} \right], 0 \left(\text{i.e., } \frac{0}{10} \right), 1 \left(\text{i.e., } \frac{10}{10} \right) \text{ and } 2 \left[\text{i.e., } \frac{20}{10} \right]$$

(4) Each value of d is multiplied by the corresponding frequency and the product is denoted by fd. In our example, the value viz., -2 is multiplied by 2, -1 is multiplied by 18; 0 is multiplied by 45; 1 is multiplied by 20; and 2 is multiplied by 6.

The sum of these products is denoted by Σfd . In the example given

$$\Sigma fd = (-2 \times 2) + (-1 \times 18) + (0 \times 45) + (1 \times 20) + (2 \times 6)$$

$$= -4 - 18 + 0 + 20 + 12 = -22 + 32 = 10$$

The total number of items is denoted as usual by Σf ,

$$\text{In our example } \Sigma f = 2 + 18 + 45 + 20 + 6 = 91$$

Now, the formula to calculate the mean value is given as follows.

$$\bar{x} = A + \frac{\Sigma fd}{\Sigma f} \times c$$

where

$$\bar{x} = \text{Mean}$$

$$A = \text{Working mean}$$

$$d = \frac{x - A}{c}$$

$$x = \text{midvalues of classes}$$

$$c = \text{width of the classes}$$

$$\Sigma fd = \text{Total sum of the products } fd$$

$$\Sigma f = \text{Total number of items}$$

In our example, $A = 25$, $\Sigma fd = 10$ and $c = 10$

$$\bar{x} = 25 + \frac{10}{91} \times 10$$

$$= 25 + \frac{100}{91}$$

$$= 25 + 1.1 \text{ (approx.)} = 26.1$$

Answer : Mean = 26.1

Example 9 :

Consider the table given below. Calculate the mean.

Class interval	Frequency
20 - 40	6
40 - 60	9
60 - 80	11
80 - 100	14
100 - 120	20

First the midvalues of the classes are calculated and they are given in the column x below. The central class the third one.

\therefore The midvalue of the third class is chosen as origin. Thus $A = 70$.
Length of each class interval = 20.

$$\therefore c = 20.$$

The deviation of all the midvalues from A are calculated and each deviation is divided by c viz., 20. The values are given under the heading d.

Each d value is multiplied by the corresponding frequency and the products are given under the heading f. The sum of these products gives Σfd .

The sum of the frequencies gives Σf .

\therefore Using $\bar{x} = A + \frac{\Sigma fd}{\Sigma f} \times c$, we can calculate the value of mean

Midvalue x	Frequency f	$\frac{x-A}{c} = d$	fd
30	6	-2	-12
50	9	-1	-9
70	11	0	0
90	14	1	14
110	20	2	40
	60		33

$$\bar{x} = A + \frac{\Sigma fd}{\Sigma f} \times c$$

$$= 70 + \frac{33}{60} \times 20$$

$$= 70 + 11 = 81$$

Answer : Mean = 81

B. Short cut method of calculating mean in the case of continuous frequency distribution with unequal class intervals.

Space for hints

Consider the following distribution :

Distribution III :

Class	Frequency
0 - 3	4
3 - 6	8
6 - 10	10
10 - 12	14

In this distribution, width of the first class is 3. This is equal to the width of the second class also. But the width of the third class is 4 and is different from the width of the previous two classes. The width of the fourth class is 2 which is different from the widths of the other three classes. Hence, it is a continuous frequency distribution with Unequal class interval.

Here, in the case of frequency distributions with unequal class intervals also, the origin A is chosen in the same way as we have explained earlier. (i.e., as we have explained in the case of continuous frequency distribution with equal class intervals.)

In the distribution III given above, there are two central classes viz., 3-6 and 6-10. Hence, A is to be chosen as the midvalue of that central class which has got greater frequency. Here the class 6-10 has greater frequency. Therefore, the midvalue of the class 6-10 viz., 8 is chosen as origin. Thus $A = 8$.

The midvalues of the classes are 1.5, 4.5, 8 and 11 respectively. Deviations of these midvalues from A are calculated. The deviations from A are -6.5 [i.e., $(1.5-8)$]-3.5 [i.e., $(4.5-8)$] 0 [i.e., $(8-8)$] 3 [i.e., $(11-8)$].

The change comes only in the definition of c. In the previous case (i.e., in the case of distributions with equal class intervals) c is defined as the width of the class intervals. But in the case of unequal class intervals, each class has a different width and hence no single value can be named as the width of the class interval for the distribution as a whole.

Check your Progress

5. What is the formula to calculate \bar{x} by shortcut method for continuous frequency distribution?

Explain the symbols used

Therefore, here we define c to be a number chosen arbitrarily, may be equal to 5 or 10 or any other number. In our example, let us take c to be equal to 5.

Now, the deviation of each midvalue x from A is divided by c . If $(x-A)$ is the deviation of a midvalue x from A , it is divided by c , $\frac{x-A}{c}$ is denoted by d .

In our example, the deviations are -6.5 , -3.5 , 0 and 3 . We have taken c to be equal to 5.

\therefore The value of d corresponding to the deviation $-6.5 = \frac{-6.5}{5} = -1.3$

The value of d corresponding to the deviation $-3.5 = \frac{-3.5}{5} = -.7$

Similarly, the other two values of d are $\frac{0}{5}$ and $\frac{3}{5}$ respectively (i.e.,) 0 and $.6$ respectively.

After the calculation of the values of d , the other things follow in the same way as in the previous case. The products of the d values and the corresponding frequencies are found out and summed up. The sum is denoted by Σfd .

All the frequencies are summed up and the sum is denoted by Σf .

Now, the formula to calculate the mean value is given as follows.

$$\bar{x} = A + \frac{\Sigma fd}{\Sigma f} \times c$$

where \bar{x} = Mean

A = Working mean value chosen

X = Midvalue of the class

c = Any arbitrary value taken

$$d = \frac{x-A}{c}$$

Σfd = sum of the products of d value and corresponding frequencies

Σf = Total number of items

Space for hints

Example 10 :

Calculate the mean value for the following distributions.

Class	Frequency
0 - 10	6
10 - 30	11
30 - 40	29
40 - 80	3
80 - 90	1

The mid values of the classes are calculated. Since there are 5 classes, we have only one central class viz., the third class. Its midvalue is 35. $\therefore A = 35$.

All the classes have unequal widths. Hence, we can choose any number as 'c'. Let us take c to be 5.

The deviation of all the midvalues from A are calculated. They are divided by c to get the values of d. Each value of d is multiplied by the corresponding frequency and the products are summed up to get Σfd .

All the frequencies are summed up to get Σf .

Midvalue x	Frequency f	$\frac{x - A}{c} = d$	fd
5	6	-6	-36
20	11	-3	-33
35	29	0	0
60	3	5	15
85	1	10	10
Total	50		-44

$$\Sigma f = 50, \Sigma fd = -44$$

$$\bar{x} = A + \frac{\Sigma fd}{\Sigma f} \times c$$

$$= 35 + \frac{-44}{50} \times 5$$

$$= 35 - 4.4 = 30.6$$

Answer : **Mean = 30.6**

2.5 Some important properties of Mean :

(i) The sum of all the deviations of given items from their mean will be equal to zero. If x is the value of one of the items and \bar{x} , the mean, then,

$$\Sigma(x - \bar{x}) = 0$$

Consider a set of the following items :

17, 16, 15, 18, 19

$$\text{Mean of the above items} = \frac{17+16+15+18+19}{5} = \frac{85}{5} = 17$$

Sum of the deviations of the given items from their mean

$$= (17-17) + (16-17) + (15-17) + (18-17) + (19-17)$$

$$= 0 - 1 - 2 + 1 + 2 = -3 + 3 = 0$$

$$\text{(i.e.,)} \Sigma(x - \bar{x}) = 0$$

(ii) Suppose the arithmetic mean of a given set of n_1 items be \bar{x}_1 and the arithmetic mean of another set of n_2 items be \bar{x}_2 . Now the mean of the combined set containing $(n_1 + n_2)$ items can be calculated by the following formula :

$$\bar{x} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2}$$

Where

n_1 = Number of items in the first set.

n_2 = Number of items in the second set.

\bar{x}_1 = Mean of the first set of items

\bar{x}_2 = Mean of the second set of items

\bar{x} = The mean of the combined set having $(n_1 + n_2)$ items.

In the same way, we can find out the mean of the combined set, when the means of more than two sets are given. For instance, if the mean of three sets are given, the combined mean viz.,

$$\bar{x} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2 + n_3 \bar{x}_3}{n_1 + n_2 + n_3}$$

If the means of 'r' sets (any given number of sets) are given, then the combined mean viz.

$$\bar{x} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2 + \dots + n_r \bar{x}_r}{n_1 + n_2 + \dots + n_r}$$

Example 11 :

60 students of a class are divided into four groups viz., the first group containing 10 students, the second group containing 25 students the third group containing 20 students and the fourth group containing 5 students. The means of the marks obtained by students belonging to the four groups are 45, 60, 58, and 50 respectively. Calculate the mean of the marks obtained by all the 60 students in the class.

n_1 = number of students in the first group = 10

\bar{x}_1 = mean of the marks obtained by students belonging to the first group = 45

n_2 = number of students in the second group = 25

\bar{x}_2 = mean of the marks obtained by students belonging to the second group = 60

In the same way,

$n_3 = 20$ $\bar{x}_3 = 58$ $n_4 = 5$ $\bar{x}_4 = 50$

Mean of the marks obtained by all the 60 students in the class

$$\bar{x} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2 + n_3 \bar{x}_3 + n_4 \bar{x}_4}{n_1 + n_2 + n_3 + n_4} = \frac{(10 \times 45) + (25 \times 60) + (20 \times 58) + (5 \times 50)}{10 + 25 + 20 + 5}$$

$$= \frac{450 + 1500 + 1160 + 250}{60} = \frac{3360}{60} = 56$$

Answer : **Arithmetic mean = 56**

3. MEDIAN

3.1 Definition :

Median is the value of the middle item of a given set of data arranged in ascending or in descending order of magnitude. Median divides the series of data given into two parts such that

1. The number of items in each part is the same.
2. The values of items in one part are greater than the value of median and the values of items in the other part are less than the value of the median.

3.2 Median for Ungrouped Data:

Let us suppose that the given set of data relates to the weight of 5 students and the weights (in lbs.) are as follows:

70, 85, 65, 100 and 90.

To locate the position of median, first of all we have to arrange the items either in ascending order of magnitude or in descending order of magnitude. Let us arrange the items given above in ascending order of magnitude as follows :

65, 70, 85, 90, 100.

Now, the third item is the middle item and its value is 85.

∴ The value of the median = 85. This value 85 divides the given set into two parts, one part containing the two numbers, 65 and 70, and the other part containing the two numbers, 90 and 100. 65 and 70 are less than 85 while 90 and 100 are greater than 85. Thus the value 85 satisfies the two properties of median which we have mentioned above.

Consider the following set of data giving the marks obtained by 6 students in an examination.

75, 60, 55, 80, 45, 70.

To find out the median let us arrange the items given above in ascending order of magnitude as follows :

45, 55, 60, 70, 75, 80.

In this illustration there are two middle items viz., the third and the fourth items.

The value of the third item = 60.

The value of the fourth item = 70

Any value taken in between 60 and 70 satisfies the two properties of the median which we have given above. For example, let us take the value 61 (between 60 and 70).

Check your Progress

6. Define Median.

This value 61 divides the given set of items into two equal parts viz., the numbers 45, 55 and 60 forming one part and the numbers 70, 75 and 80 forming the other part.

Space for hints

Each part has three items.

All the items in one part viz., 45, 55 and 60 are less than 61, all the items in the other part viz., 70, 75 and 80 are greater than 61.

Hence, 61 may be taken as the value of median.

In the same way, some other number in between 60 and 70 may also be taken as the median.

But it is usual to take the value midway between the central values as the value of the median.

In our example, the point midway between 60 and 70 is taken to be the median.

The point midway between 60 and 70 is got by adding 60 and 70 and dividing the sum by 2.

$$\therefore \text{In our example, the value of the median} = \frac{60+70}{2} = \frac{130}{2} = 65$$

Hence, to find out the median the following rule is adopted.

First arrange the given set of items in ascending (or in descending) order of magnitude.

If n is the number of items given find out the value of the item $\left\{\frac{n+1}{2}\right\}$.

The value of the item $\left\{\frac{n+1}{2}\right\}$ is the value of median.

For example, if there are 7 items given, then the value of $\left\{\frac{7+1}{2}\right\}$ th item is the median. That is, the value of the 4th item is the median. If, for instance, there are 8 items given, then the value of median

$$= \text{the value of } \left\{\frac{8+1}{2}\right\} \text{th item}$$

$$= \text{the value of } \left\{\frac{9}{2}\right\} \text{th item}$$

$$= \text{value of 4.5th item}$$

$$= \frac{\text{value of 4th item} + \text{value of 5th item}}{2}$$

The value of median is also expressed in the same units in which the values of the given items are expressed. For example, the given set of data

gives the weights of certain number of persons in pounds, the median is also given in pounds.

Consider the following worked out examples.

Example 1 :

Locate the median income from the following data of income (in Rs.) received by 11 workers in a Company.

50, 75, 60, 55, 52, 68, 72, 57, 53, 64, 69

Let us arrange the items given above in ascending order of magnitude as follows:

50, 52, 53, 55, 57, 60, 64, 68, 69, 72, 75

n = Total number of items given = 11

$$\begin{aligned}\therefore \text{Median} &= \text{The value of the item } \left\{ \frac{11+1}{2} \right\} \\ &= \text{The value of the item } \left\{ \frac{12}{2} \right\} \\ &= \text{The value of the 6th item} = \text{Rs. 60}\end{aligned}$$

Answer : Median = Rs. 60.

Example 2 :

The marks obtained by 10 students are given below, Locate the median

93, 78, 85, 90, 60, 30, 45, 50, 70, 49

Let us form an array of the marks given as follows :

30, 45, 49, 50, 60, 70, 78, 85, 90, 93

n = Total number of items given = 10

$$\begin{aligned}\therefore \text{Median} &= \text{The value of the item } \left\{ \frac{10+1}{2} \right\} \\ &= \text{The value of the item } \left\{ \frac{11}{2} \right\} \\ &= \text{The value of the item } 5.5 \\ &= \frac{\text{Value of the 5th item} + \text{Value of the 6th item}}{2} \\ &= \frac{60+70}{2} = \frac{130}{2} = 65\end{aligned}$$

Answer : Median = 65

Example 3 :

Space for hints

The monthly income of eight families in (rupees) are given below.
Calculate the median

Family	Income (in Rs.)
A	30
B	70
C	10
D	75
E	500
F	8
G	42
H	250

Let us form an array of the incomes given as follows :

8, 10, 30, 42, 70, 75, 250, 500.

Total number of items = $n = 8$

\therefore Median Income = The value of the item $\left\{ \frac{8+1}{2} \right\}$

= The value of the item 4.5

= $\frac{\text{Value of the 4th item} + \text{Value of the 5th item}}{2}$

= Rs. $\frac{42+70}{2}$ = Rs. $\frac{112}{2}$ = Rs. 56

Answer : Median = 56

3.3 Median for Discrete Frequency Distribution :

Suppose we are given a discrete frequency distribution as follows.

Value of the item	Frequency
55	8
58	10
59	7
62	4
Total	29

To calculate the median, first of all we find out the total frequency which is denoted by N.

In our example, total frequency is 29 and is denoted by N

$$\therefore N = 29$$

Now median is given by the following formula :

$$\text{Median} = \text{Value of the item } \left\{ \frac{N+1}{2} \right\}$$

In our example,

$$\text{Median} = \text{Value of the item } \left\{ \frac{29+1}{2} \right\} = \text{Value of the item 15}$$

To find out the value of the item $\frac{N+1}{2}$ we cumulate the given frequencies. The cumulation is done as follows :

With each frequency given, we add up all the preceding frequencies. Since the first frequency does not have any preceding frequency, we retain the first frequency as it is; with the second frequency we add up the first frequency; with the third frequency we add up the first two frequencies and so on.

In our example, the first frequency 8 is kept as it is. That is, 8 is the first cumulative frequency.

The second frequency is 10 and with it the preceding frequency 8 is added. The sum is 18 and it is the second cumulative frequency.

The third frequency is 7 and with it the two preceding frequencies 8, and 10 are added. The sum is 25 and it is the third cumulative frequency.

The fourth frequency is 4 and with it the three preceding frequencies 8, 10 and 7 are added. The sum is 29 and it is the fourth cumulative frequency.

Now we give the cumulative frequencies in the form of a table as follows :

Value of the item	Cumulative frequency
55	8
58	18
59	25
62	29

From the above table we get the following information.

Space for hints

In the array of the given set of data, value of each of the first 8 items is equal to 55; after the 8th item upto the 18th item, value of each item is equal to 58; after the 18th item upto the 25th item, value of each item is equal to 59; after the 25th item upto the 29th item value of each item is equal to 62.

We have to find out the value of the item 15 using the above information. The item 15 comes after the 8th item but before the 18th item. Hence, its value is equal to 58.

$\therefore \text{Median} = 58$

Now we give the step by step procedure to calculate the median of a discrete frequency distribution as follows :

- 1) Find out the total frequency and denote it by N .
- 2) Find out the value of $\frac{N+1}{2}$
- 3) With each frequency add all the preceding frequencies and thus find out the cumulative frequencies.
- 4) Now, in the cumulative frequency column find out two cumulative frequencies with in which $\frac{N+1}{2}$ falls;
- 5) Of the two cumulative frequencies found out above, consider the bigger one. The value of the item corresponding to this bigger cumulative frequency gives the value of the item $\frac{N+1}{2}$.

In the case of discrete frequency distribution

Median = Value of the item $\frac{N+1}{2}$

\therefore The value of the item $\frac{N+1}{2}$ obtained above gives us the required value of median.

Example 4 :

Value of the item (x)	Frequency (f)
4	2
5	5
6	8
7	9
8	12
9	11
10	13
Total	60

Find out the median for the distribution given above.

$$N = \text{Total frequency} = 60$$

$$\therefore \frac{N+1}{2} = \frac{60+1}{2} = \frac{61}{2} = 30.5$$

We cumulate the frequencies and give them as follows :

Value of the item (x)	Frequency (f)
4	2
5	7
6	15
7	24
8	36
9	47
10	60

In the above cumulative frequency column, 24 and 36 are the two cumulative frequencies with in which $\frac{N+1}{2}$ viz. 30.5 falls. Now, the value of the item corresponding to the bigger one of the two cumulative frequencies viz., 36 is 8. And it is the value of the item $\frac{N+1}{2}$. That is, 8 is the value of the item 30.5 in our example.

$$\therefore \text{Median} = \text{value of the item } \frac{N+1}{2} = \text{value of the item } 30.5 = 8$$

Example 5 :

Space for hints

Calculate the Median from the following data :

Value	Frequency
15	3
20	21
23	46
25	52
35	10
40	6
47	1
Total	139

 $N = \text{Total frequency} = 139.$

$$\frac{N+1}{2} = \frac{139+1}{2} = \frac{140}{2} = 70$$

Value	Cumulative Frequency (cf)
15	3
20	24
23	70
25	122
35	132
40	138
47	139

Here $\frac{N+1}{2}$ is equal to the third cumulative frequency, 70. Hence value of the item corresponding to the cumulative frequency 70 is the value of item $\frac{N+1}{2}$. In the above table, 23 is the value corresponding the cumulative frequency 70.

$$\begin{aligned}\therefore \text{Median} &= \text{value of the item } \frac{N+1}{2} \\ &= \text{value of the item 70} = 23\end{aligned}$$

3.4 Median for Continuous Frequency Distribution

In the case of continuous frequency distribution, median is defined as the value of the item $\frac{N}{2}$. Here we meet with one difficulty. The value of median lies in a 'class interval'. To get a definite figure, the value of median is to be estimated. We have explained below the step by step procedure to estimate the value of median from a continuous frequency distribution :

- 1) To calculate median we need true class intervals. So, first of all we must see whether the intervals in the given distribution are true class intervals or not. If not, we must convert them into true class intervals.
- 2) We find out the total frequency and denote it by N .
- 3) We find out half of the total frequency i.e., $\frac{N}{2}$.
- 4) We find out the less than cumulative frequencies from the given simple frequencies and give them under the heading 'cumulative frequency'.

- 5) We find out the two cumulative frequencies with in which $\frac{N}{2}$ falls.
- 6) Of the two cumulative frequencies, we consider the second one and the corresponding class interval. This is the class interval in which the value of the item $\frac{N}{2}$ falls. By definition,

$$\text{Median} = \text{value of the item } \frac{N}{2}$$

\therefore The class in which the value of the item $\frac{N}{2}$ falls is the same as the class in which the value of median falls.

Hence this class is called the median class.

- 7) True lower limit of the median class is denoted by l .
- 8) Cumulative frequency of the class immediately preceding median class is found out and is denoted by m .
- 9) Frequency of the median class is denoted by f .

10) Length of the median class is found out and is denoted by c . Now, the formula to get the value of median is as follows.

Space for hints

$$\text{Median} = l + \frac{\frac{N}{2} - m}{f} \times c$$

Let us consider the following example.

Example 6 :

Calculate the median from the following data :

Class	Frequency
0 - 5	4
5 - 10	8
10 - 15	10
15 - 20	14
20 - 25	16
25 - 30	22

In the given frequency table the intervals are true class intervals. Therefore, there is no necessity for the conversion of given intervals into true class intervals.

$$N = \text{total frequency} = 74$$

$$\frac{N}{2} = \frac{74}{2} = 37$$

We find out the less than cumulative frequencies and give them as follows :

Class	Frequency	Cumulative Frequency
0 - 5	4	4
5 - 10	8	12
10 - 15	10	22
15 - 20	14	36
20 - 25	16	52
25 - 30	22	74

36 and 52 are two cumulative frequencies with in which $\frac{N}{2}$ falls. Of the two cumulative frequencies 52 is the bigger cumulative frequency and the corresponding class interval is 20-25. In this class only the value of the item $\frac{N}{2}$ falls. That is, in the class interval 20-25 the value of median falls.

\therefore 20-25 is called the median class.

Now the true lower limit of the median class is 20 and is denoted by l .

$$\therefore l = 20$$

Cumulative frequency of the class immediately preceding median class is 36 and is denoted by m .

$$\therefore m = 36$$

Frequency of the median class is 16 and is denoted f . $\therefore f = 16$

Length of the median class is $25 - 20 = 5$ and is denoted by c . $\therefore c = 5$.

$$\begin{aligned} \text{Now, Median} &= l + \frac{\frac{N}{2} - m}{f} \times c \\ &= 20 + \frac{37 - 36}{16} \times 5 \\ &= 20 + \frac{1}{16} \times 5 \\ &= 20 + .31 = 20.31 \end{aligned}$$

Answer : Median = 20.31

Example 7 :

Calculate the Median.

Weight of students in lbs	Frequency
0 - 10	32
10 - 20	65
20 - 30	100
30 - 40	184
40 - 50	288
50 - 60	167
60 - 70	98
Total	934

The given class intervals are true class intervals.

$N = \text{Total frequency} = 934$

$\therefore \frac{N}{2} = \frac{934}{2} = 467$

Class Interval	Frequency	Cumulative Frequency
0 - 10	32	32
10 - 20	65	97
20 - 30	100	197
30 - 40	184	381
40 - 50	288	669
50 - 60	167	836
60 - 70	98	934

$\frac{N}{2}$ is viz., 467 lies between the cumulative frequencies 381 and 669. Therefore, the class corresponding to the cumulative frequency 669. viz., 40–50 is the median class.

$l = \text{True lower limit of the median class} = 40$

$m = \text{cumulative frequency of the class just above the median class.}$

$= \text{cumulative frequency of the class '30–40'} = 381$

$f = \text{frequency of the median class} = 288$

$c = \text{magnitude of the median class} = 50 - 40 = 10$

$\therefore \text{Median} = l + \frac{\frac{N}{2} - m}{f} \times c$

$= 40 + \frac{467 - 381}{288} \times 10$

$$= 40 + \frac{86}{144} \times 5 = 40 + \frac{215}{72}$$

$$= 40 + 3 \text{ (approx.)} = 43$$

Example 8 :

Calculate the median.

Class Interval	Frequency
10 - 19	52
20 - 29	61
30 - 39	190
40 - 49	67
50 - 59	45
60 - 69	40

The class intervals in the distribution given above are not true class intervals. Therefore, we have to convert them into true class intervals.

We have converted the given class intervals in to true class intervals and given them in the table below. Also we have given the cumulative frequencies.

Class Interval	Frequency	Cumulative Frequency
9.5 - 19.5	52	52
19.5 - 29.5	61	113
29.5 - 39.5	190	303
39.5 - 49.5	67	370
49.5 - 59.5	45	415
59.5 - 69.5	40	455
Total	455	

$$N = 455$$

$$\frac{N}{2} = \frac{455}{2} = 227.5$$

$\frac{N}{2}$ is viz., 227.5 lies between the cumulative frequencies 113 and 303.

So, the class corresponding to the cumulative frequency 303 is the median class. That is, "29.5–39.5" is the median class.

l = True lower limit of the median class = 29.5

f = frequency of the median class = 190

c = magnitude of the median class = $39.5 - 29.5 = 10$

m = cumulative frequency of the class just above the median class.

= cumulative frequency of the class '19.5–29.5' = 113

$$\begin{aligned}\therefore \text{Median} &= l + \frac{\frac{N}{2} - m}{f} \times c \\ &= 29.5 + \frac{227.5 - 113}{190} \times 10\end{aligned}$$

$$= 29.5 + \frac{114.5}{19} = 29.5 + 6 = 35.5$$

Answer :

Median = 35.5

Example 9 :

Calculate the median of the following data.

Monthly Earning Rs.	No. of workers
28 - 32	30
33 - 37	40
38 - 42	50
43 - 47	150
48 - 52	300
53 - 57	39
Total	609

The class intervals of the distribution given above are not true class intervals. To calculate the median, the class intervals must be true class

intervals. Therefore we convert the class intervals given above in to true class intervals and give them in the following table.

Monthly wage Rs.	Frequency	Cumulative Frequency
27.5 - 32.9	30	30
32.5 - 37.5	40	70
37.5 - 42.5	50	120
42.5 - 47.5	150	270
47.5 - 52.5	300	570
52.5 - 57.5	39	609
Total	609	

$$N = 609$$

$$\frac{N}{2} = \frac{609}{2} = 304.5$$

$\frac{N}{2}$ is 304.5 and it lies between the cumulative frequencies 270 and 570.

The class corresponding to the cumulative frequency 570 viz., "47.5-52.5" is the median class.

$$\therefore l = 47.5 \quad f = 300 \quad c = 52.5 - 47.5 = 5 \quad m = 270$$

$$\therefore \text{Median} = l + \frac{\frac{N}{2} - m}{f} \times c$$

$$= 47.5 + \frac{304.5 - 270}{300} \times 5$$

$$= 47.5 + \frac{34.5}{300} \times 5$$

$$= 47.5 + \frac{172.5}{300}$$

$$= 47.5 + .575 = 48.075$$

Answer : Median = 48.075

3.5 Calculation of Median when the Midvalues of the Class Intervals are given :

Space for hints

Sometimes the class intervals of the distribution may not be given. Instead of that, the midvalues of the class intervals may be given. In such cases, to calculate the median, the class intervals should first be found out from the midvalues. Then the median class is found out as we have explained above and median is calculated.

Consider the following example.

Example 10 :

Calculate the median from the following data.

Midvalue	Frequency
15	10
25	25
35	36
45	46
55	52
65	31
75	28
85	22

First the class intervals are found out from the midvalues given. We have given the class intervals in the table below. We have also given the cumulative frequencies.

Class	Frequency	Cumulative Frequency
10 - 20	10	10
20 - 30	25	35
30 - 40	36	71
40 - 50	46	117
50 - 60	52	169
60 - 70	31	200
70 - 80	28	228
80 - 90	22	250

Space for hints

$$N = \text{Total frequency} = 250$$

$$\frac{N}{2} = \frac{250}{2} = 125$$

Using cumulative frequency column we get the median class to be 50–60

$$\therefore l = 50 \quad m = 117 \quad f = 52 \quad c = 60 - 50 = 10$$

$$\therefore \text{Median} = l + \frac{\frac{N}{2} - m}{f} \times c$$

$$= 50 + \frac{125 - 117}{52} \times 10 = 50 + \frac{8}{52} \times 10 = 50 + \frac{20}{13}$$

$$= 50 + 1.54 = 51.54$$

Answer : Median = 51.54

Example 11 :

Calculate the median from the following frequency distribution.

Midvalue	frequency
7.5	12
22.5	21
37.5	23
52.5	34
67.5	10

First we find out the class intervals. We have given them in the following table. We have given the cumulative frequencies also in the same table.

Class Interval	Frequency	Cumulative Frequency
0 - 15	12	12
15 - 30	21	33
30 - 45	23	56
45 - 60	34	90
60 - 75	10	100

$N = \text{Total frequency} = 100$

Space for hints

$$\frac{N}{2} = \frac{100}{2} = 50$$

From the above table, we get that '30–45' is the median class

$$\therefore l = 30 \quad m = 33 \quad f = 23 \quad c = 15$$

$$\therefore \text{Median} = l + \frac{\frac{N}{2} - m}{f} \times c$$

$$= 30 + \frac{50 - 33}{23} \times 15 = 30 + \frac{17}{23} \times 15$$

$$= 30 + \frac{255}{23} = 30 + 11.1 = 41.1$$

Answer : Median = 41.1

3.6 Calculation of Median when Cumulative Frequency Distribution (either less than or more than cumulative frequency distribution) is given :

Sometimes we may be given either a less than or a more than cumulative frequency distribution and we may be asked to calculate the median. In such cases, first of all we must get the simple frequency distribution from the given cumulative frequency distribution. Then as usual median is calculated using the formula.

$$\therefore \text{Median} = l + \frac{\frac{N}{2} - m}{f} \times c$$

Example 12 :

Calculate the median from the following distribution.

Space for hints

Marks less than	Frequency
10	5
20	9
30	17
40	29
50	45
60	60

We are given a less than cumulative frequency distribution. To calculate median we need class intervals. So, we convert the given less than cumulative frequency distribution in to a simple frequency distribution and give it as follows :

Class Interval	Frequency	Cumulative Frequency
0 – 10	5	5
10 – 20	4	9
20 – 30	8	17
30 – 40	12	29
40 – 50	16	45
50 – 60	15	60

Now we calculate median as usual

$$N = \text{Total frequency} = 60$$

$$\frac{N}{2} = \frac{60}{2} = 30$$

From the cumulative frequency column, we get the class 40-50 as the median class.

$$\therefore l = 40 \quad m = 29 \quad f = 16 \quad c = 10$$

$$\therefore \text{Median} = l + \frac{\frac{N}{2} - m}{f} \times c$$

$$= 40 + \frac{30-29}{16} \times 10$$

$$= 40 + \frac{1}{16} \times 10$$

$$= 40 + .625$$

$$= 40.625$$

Answer : Median = 40.625

4. QUARTILES

We have seen in the previous topic that the median divides an array of the given items into two equal parts. The values of items in one part are greater than the median value, and the values of items in the other part, less than median value. In order to have a better idea about the composition of a given series, it may be necessary to divide it in to four, ten or hundred equal parts. Just as one item divides a series in to two equal parts, three items would divide it into four equal parts, nine items into ten equal parts and ninety nine items into hundred equal parts. The values of items are respectively known as quartiles, deciles and percentiles. As quartiles alone are included in your syllabus, we give below the computation of quartiles only.

4.1 Quartiles - Meaning and Definition :

When a given distribution of items arranged in ascending order is divided into four equal parts, we get three dividing positions. The values of items in these three dividing positions are called the 'quartiles'.

The value of the item in the first dividing position will be such that its value is greater than the values of one fourth of the given items. Hence, the value of the item in the first dividing position is called the first quartile or the lower quartile.

The value of the item in the second dividing position will be such that its value is greater than the value of two fourths (i.e.,) half of the given items. It is to be noted that, this value is the same as the value of the median. The value of the item in the second dividing position is called the second quartile. Hence, median gets another name, viz., 'the second quartile'.

The value of the item in the third dividing position will be such that its value is greater than three fourths of the given items. This value is called the third quartile or the upper quartile.

We have given below an account of the lower and upper quartiles.

(i) Lower Quartile (or the first quartile) - Computation procedure :

Lower quartile is defined as the value which is such that one fourth of the given items have their values below it, while three fourth of the given item have their values above it. Lower quartile is usually denoted by Q_1 .

Lower quartile is also called the 'first quartile'.

(A) Calculation of Lower Quartile (Viz., Q_1) From ungrouped Data :

To calculate the lower quartile,

- 1) The given items are arranged in ascending order of magnitude.
- 2) Total number of items given is denoted by n . Now the lower quartile is given by the following formula,

$$Q_1 = \text{Value of the item} \left(\frac{n+1}{4} \right)$$

Suppose, we have to find out the lower quartile of the following numbers :

35, 80, 50, 60, 48, 63, 65

First of all the given items are arranged in ascending order of magnitude as follows :

35, 48, 50, 60, 63, 65, 80

$n =$ Total number of items given

$$\therefore Q_1 = \text{Value of the item} \left(\frac{n+1}{4} \right)$$

Check your Progress

7. What are Quartiles? How many Quartiles are there?

$$\left(\frac{1+n}{4}\right) = \text{Value of the item } \left(\frac{7+1}{4}\right)$$

$$= \text{Value of the item } \left(\frac{8}{4}\right)$$

$$= \text{Value of the item (2)} = 48$$

$$\therefore \text{Lower quartile} = 48$$

In the illustration given above $(n+1)$ viz., 8 is exactly divisible by 4 and the quotient we get is 2, a whole number. So, we easily find out the value of Q_1 which is equal to the value of item 2 in the array viz., 48.

But in some cases $(n+1)$ may not be exactly divisible by 4. In such cases, quotient is a whole number followed by any one of the fractions, .25, .5, and .75. In such cases where we get the value of $\left(\frac{n+1}{4}\right)$ to be a composite number we calculate the value Q_1 as follows

1. We note down the value of the item denoted by the whole number.
2. We find out the difference between the value of the item denoted by the whole number and the value of the item which is immediately succeeding the item denoted by the whole number.
3. We multiply this difference by the fractional part.
4. We add together the value of an item denoted by the whole number and the product that we get by multiplying 'the differences by the fraction'. This gives the value of Q_1 (i.e.,) lower quartile.

Wherever we have to find out the value of an item denoted by a composite number, we adopt the same procedure as explained above.

Consider the following examples :

Example 1 :

Suppose the following numbers represent the given items. Calculate their lower quartile.

35, 48, 50, 60, 63, 65.

n = Total number of items given.

$$Q_1 = \text{Value of the item } \left(\frac{n+1}{4} \right)$$

$$= \text{Value of the item } \left[\frac{7}{4} \right] = \text{value of the item 1.75}$$

We calculate the value of the item 1.75 as given below :

$$\begin{aligned} \text{Value of the item 1.75} &= (\text{value of the item 1}) + \frac{3}{4}(\text{value of the item 2} - \text{value of item 1}) \\ &= 35 + \frac{3}{4}(48 - 35) = 35 + \frac{39}{4} = 35 + 9.75 = 44.75 \end{aligned}$$

Lower quartile = 44.75

Example 2 :

Suppose the following numbers are the given items. Calculate the lower quartile.

35, 48, 50, 60, 63, 65, 80, 85.

Here

n = Total number of items given.

$$\therefore Q_1 = \text{Value of the item } \left(\frac{n+1}{4} \right)$$

$$= \text{Value of the item } \left(\frac{8+1}{4} \right)$$

$$= \text{value of the item } \left(\frac{9}{4} \right)$$

$$= \text{value of the item 2.25}$$

We calculate the value of the item 2.25 as follows :

$$\begin{aligned} \text{Value of the item 2.25} &= \text{value of the item 2} + \frac{1}{4}(\text{value of the item 3} - \text{value of the item 2}) \\ &= 48 + \frac{1}{4}(50 - 48) = 48 + \frac{1}{4} \times 2 = 48 + 0.5 = 48.5 \end{aligned}$$

Lower quartile = 48.5

Example 3 :

Suppose the following numbers represent the given set of items. Calculate the lower quartile.

35, 48, 50, 60, 68, 65, 80, 85, 88

n = Total number of items 9

$$Q_1 = \text{Value of the item } \left(\frac{n+1}{4} \right)$$

$$= \text{Value of the item } \left(\frac{9+1}{4} \right)$$

$$= \text{value of the item } \left(\frac{10}{4} \right)$$

$$= \text{value of the item 2.5}$$

The value of the item 2.5 is calculated as follows :

$$\text{Value of the item 2.5} = \text{value of the item 2} + \frac{1}{2}(\text{value of the item 3} - \text{value of item 2})$$

$$= 48 + \frac{1}{2}(50 - 48)$$

$$= 48 + \left(\frac{1}{2} \times 2 \right) = 48 + 1 = 49$$

Lower quartile = 49

B) Calculation of lower quartile from discrete frequency distribution :

To calculate the lower quartile.

- 1) The given frequency distribution is converted in to cumulative (less than cumulative) frequency distribution.
- 2) Total sum of all the frequencies is found out. The sum is denoted by N and the value of $\frac{N+1}{4}$ is found out.
- 3) Lower quartile is calculated using the formula,

$$Q_1 = \text{value of the item } \left(\frac{N+1}{4} \right)$$

Where Q_1 denotes the lower quartile.

Consider the following discrete frequency distribution :

Space for hints

Value of item	Frequency
55	8
65	10
75	16
85	14
95	10
105	5
115	2

The given frequencies are converted into cumulative frequencies and given as follows :

Value of the item	Frequency	Cumulative Frequency
55	8	8
65	10	18
75	16	34
85	14	48
95	10	58
105	5	63
115	2	65
Total	65	-

$$N = \text{Total frequency} = 65$$

$$\frac{N+1}{4} = \frac{65+1}{4} = \frac{66}{4} = 16.5$$

$$Q_1 = \text{Value of the item} \left[\frac{N+1}{4} \right]$$

$$= \text{value of the item } 16.5$$

It is to be noted from the cumulative frequency column of the above table, that all the items beyond the item 8 and upto the item 18 are having their values equal to 65. The item 16.5 is in between the items 8 and 18. Hence the value of the item 16.5 is also equal to 65.

$$\therefore Q_1 = 65$$

(C) Calculation of lower quartile from continuous frequency distribution :

Space for hints

- 1) The frequencies in the given distribution are converted into cumulative frequencies.
- 2) All the given frequencies are summed up and the sum is denoted by N.
The value of $\frac{N}{4}$ is found out.
 Q_1 is defined as
 $Q_1 = \text{Value of the item } \frac{N}{4}$
- 3) The class containing the value of Q_1 is found out. This class is called the Q_1 class.
- 4) Now lower quartile is calculated using the formula.

$$Q_1 = l + \frac{\frac{N}{4} - m}{f} \times c$$

Where Q_1 denotes lower quartile

l denotes the true lower limit of Q_1 class.

N denotes the total frequency

m denotes the cumulative frequency of the class immediately preceding the Q_1 class.

f denotes the frequency of the Q_1 class.

c denotes the magnitude or length of the Q_1 class

Consider the following continuous frequency distribution

Class	Frequency
0 - 3	4
3 - 6	8
6 - 9	10
9 - 12	14
12 - 15	7

The given frequencies are converted into cumulative frequencies and given as follows :

Class	Frequency	Cumulative Frequency
0 - 3	4	4
3 - 6	8	12
6 - 9	10	22
9 - 12	14	36
12 - 15	7	43
Total	43	

$$N = \text{Total frequency} = 43$$

$$\frac{N}{4} = \frac{43}{4} = 10.75$$

$$Q_1 = \text{Value of the item} \left[\frac{N}{4} \right]$$

From the cumulative frequency column in the table given above, we come to know that all the items after item 4 and upto the item 12 are having their values in the interval '3-6'.

The item 10.75 is in between the items 4 and 12. Hence, its value is also in the interval '3-6'.

(i.e.,) the value of Q_1 lies in the interval '3-6'.

Q_1 class is '3-6'

$$l = \text{true lower limit of the } Q_1 \text{ class} = 3$$

$$m = \text{cumulative frequency of the class preceding the } Q_1 \text{ class.}$$

$$= \text{cumulative frequency of the class '0-3'} = 4$$

$$f = \text{frequency of the } Q_1 \text{ class} = 8$$

$$c = \text{magnitude of the } Q_1 \text{ class} = 6-3 = 3$$

$$Q_1 = l + \frac{\frac{N}{4} - m}{f} \times c$$

$$= 3 + \frac{10.75 - 4}{8} \times 3$$

$$= 3 + \frac{6.75}{8} \times 3$$

$$= 3 + \frac{20.25}{8}$$

$$= 3 + 2.53 = 5.53$$

$$\therefore \text{Lower quartile} = 5.53$$

(ii) Upper quartile (or the third quartile) computation procedure :

Upper quartile is defined as that value which is such that three fourths of the given items have their values below it and one fourth of the given items have their values above it. Upper quartile is usually denoted by Q_3 . Upper Quartile is also known as the 'third quartile'.

(A) Calculation of upper quartile from ungrouped data :

To calculate the upper quartile we adopt the following method :

1. The given items are arranged in ascending order of magnitude.
2. Total number of items given is found out and is denoted by n .
3. Upper quartile is found out using the formula.

$$Q_3 = \text{value of the item } \frac{3(n+1)}{4}$$

Where Q_3 denotes the upper quartile.

Consider the following example.

Example 5 :

Calculate the upper quartile of the following items.

28, 32, 27, 45, 43, 52.

The given items are arranged in ascending order of magnitude as follows :

27, 28, 32, 43, 45, 52

$$n = \text{Total number of items given} = 6$$

$$Q_3 = \text{value of the item } \frac{3(n+1)}{4}$$

$$= \text{value of the item } \frac{3(6+1)}{4}$$

$$= \text{value of the item } \frac{3 \times 7}{4}$$

$$= \text{value of the item } 5.25$$

Just as in the case of the lower quartile here also we find out value of the item 5.25 as follows :

$$\text{Value of the item } 5.25 = \text{value of the item } 5 + \frac{1}{4}(\text{value of the item } 6 - \text{value of item } 5)$$

$$= 45 + \frac{1}{4}(52 - 45) = 45 + \frac{7}{4}$$

$$= 45 + 1.75 = 46.75$$

$$\text{Upper quartile} = 46.75$$

Example 6 :

Calculate the upper quartile of the following items.

27, 28, 32, 38, 43, 45, 52.

Here, the items are already in an ascending order of magnitude.

$$n = \text{Total number of items given} = 7$$

$$Q_3 = \text{value of the item } \frac{3(n+1)}{4}$$

$$= \text{value of the item } \frac{3(7+1)}{4}$$

$$= \text{value of the item } \frac{3 \times 8}{4}$$

$$= \text{value of the item } 6 = 45$$

$$\text{Upper quartile} = 45$$

B) Calculation of upper quartile from discrete frequency distribution :

Upper quartile is calculated using the following method :

1. Cumulative frequencies (less than cumulative frequencies) of the given frequencies are calculated.
2. Sum of all the frequencies is found out and is denoted by N.

3. Upper quartile is calculated using the formula.

Space for hints

$$Q_3 = \text{value of the item } \frac{3(N+1)}{4}$$

Consider the following example.

Example 7 :

Calculate the third quartile of the following distribution.

Value of item	Frequency
55	8
65	10
75	16
85	14
95	10
105	5
115	2

Cumulative frequencies are calculated and given in the following table :

Value of the item	Frequency	Cumulative Frequency
55	8	8
65	10	18
75	16	34
85	14	48
95	10	58
105	5	63
115	2	65
Total	65	-

$$N = \text{Total frequency} = 65$$

$$\frac{3(N+1)}{4} = \frac{3(65+1)}{4} = \frac{3 \times 66}{4} = 49.5$$

$$Q_3 = \text{Value of the item } \frac{3(N+1)}{4}$$

$$= \text{value of the item } 49.5$$

We come to know from the cumulative frequency column of the above table, that all items beyond the item 48 and upto the item 58 are having their values equal to 95.

The item 49.5 is in between the items 48 and 58.

∴ Its value is also equal to 95.

Therefore, $Q_3 = 95$.

Third or upper quartile = 95.

(C) Calculation of upper quartile from continuous frequency distribution :

To calculate the upper quartile we adopt the following procedure :

- 1) All the frequencies are summed up and the sum is denoted by N.
- 2) The upper quartile is defined as

$$Q_3 = \text{value of the item } \left(\frac{3N}{4} \right)$$

The class in which the value of the upper quartile falls is found out. This class is called Q_3 class.

To find out Q_3 class, the given frequencies are cumulated first. Using the cumulative frequencies the class in which the item, $\left(\frac{3N}{4} \right)$ falls is found out. Thus Q_3 class is found out.

- 3) Now the upper quartile is calculated using the formula,

$$Q_3 = l + \frac{\frac{3N}{4} - m}{f} \times c$$

where Q_3 denotes the upper quartile

l denotes the true lower limit of the Q_3 class.

m denotes the cumulative frequency of the class immediately preceding the Q_3 class.

f denotes the frequency of the Q_3 class.

c denotes the magnitude or length of the Q_3 class.

Consider the following example.

Example 8 :

Calculate the upper quartile of the following distribution

Class	Frequency
0 - 3	4
3 - 6	8
6 - 9	10
9 - 12	14
12 - 15	7

The cumulative frequencies are calculated and given in the following table :

Class	Frequency	Cumulative Frequency
0 - 3	4	4
3 - 6	8	12
6 - 9	10	22
9 - 12	14	36
12 - 15	7	43
Total	43	

$$N = \text{Total frequency} = 43$$

$$\frac{3N}{4} = \frac{3 \times 43}{4} = \frac{129}{4} = 32.25$$

$$Q_3 = \text{Value of the item } \left[\frac{3N}{4} \right]$$

$$= \text{Value of the item } 32.25.$$

We come to know from the cumulative frequency column of the above table that all items beyond the item 22 and up to the item 36 have their values lying in the interval '9-12'.

The item 32.25 is in between the items 22 and 36.

∴ The item 32.25 has its value in the interval '9-12'

(i.e.,) The upper quartile lies in the interval '9-12'

(i.e.,) '9-12' is the Q_3 class.

$$l = \text{true lower limit of the } Q_3 \text{ class} = 9$$

$$m = \text{cumulative frequency of the class immediately preceding the } Q_3 \text{ class.}$$

$$= \text{cumulative frequency of the class '0-9'} = 22$$

$$f = \text{frequency of the } Q_3 \text{ class} = 14$$

$$c = \text{magnitude of the } Q_3 \text{ class} = 12 - 9 = 3.$$

$$Q_3 = l + \frac{\frac{3N}{4} - m}{f} \times c$$

$$= 9 + \frac{32.25 - 22}{14} \times 3 = 9 + \frac{10.25}{14} \times 3$$

$$= 9 + \frac{30.75}{14} = 9 + 2.196$$

$$= 11.196$$

$$\text{Upper quartile} = 11.196$$

5. MODE

Space for hints

5.1 Definition :

The mode is defined as that value which occurs most frequently. Mode represents the value which occurs most in a group; the value which is in fact the fashion (1a mode). When we speak of the 'average student', 'the most common wage', 'the common man', or 'the typical farm' and the like we are referring to the mode not being aware of it. If it is said that the 'average' size of shoes sold is the size nine, it means that most of the shoes sold are of size nine. Other sizes of shoes sold are lesser in number compared to the size nine. In this case also we are talking about mode (viz., the average size of shoes) but not being aware of it.

5.2 Calculation of mode from ungrouped data :

Location of mode in the case of ungrouped data does not present any difficulty. The given set of items are arranged in order of magnitude. From the array formed, the value which occurs the greatest number of times is found out. This value will be the mode of the given set of items. For example, consider the following set of items.

8, 10, 9, 17, 10, 19, 15, 10, 12, 19

Let us arrange these items in ascending order of magnitude

8, 9, 10, 10, 10, 12, 15, 17, 19, 19.

The value 10 occurs three times; the value 19 occurs two times and all the other values occur once. Thus, the value 10 occurs greater number of times than others. Hence, the mode is 10.

A slight change in the set of values given may change the value of the mode considerably. If in the above example, one 10 is replaced by 19, then the array becomes,

8, 9, 10, 10, 12, 15, 17, 19, 19, 19.

Here 19 occurs three times; 10 occurs two times and all the other values occur once. 19 occurs greater number of times than others and it is the modal value.

Check your
Progress

8. Define Mode.

Hence, a change of one item has altered the mode from 10 to 19. Thus the modal value is highly unstable.

Consider the first set given above viz.

8, 9, 10, 10, 10, 12, 15, 17, 19, 19

Here the modal value is 10 and is unique. Such a set having only one value as the modal value is called 'unimodal'.

Consider the following set of items :

8, 9, 10, 10, 10, 12, 15, 17, 19, 19, 19

Both the values 10 and 19 occur three times and all the other values occur only once. Hence, both 10 and 19 can be taken as the mode. Thus, the given set has two modal values.

A set having two modal values is called "bi-modal".

In the same way, if a set has three modal values, then it is called 'tri-modal'.

If a set has more than three modal values, then it is called 'multi-modal'.

Consider the following set of items :

5, 8, 9, 11, 15.

In this set all the values occur once and no value occurs greater number of times than others. Therefore, the modal value cannot be determined. In such a case the set is said to have 'no-mode'.

If we have a set of items which is of the above form but the number of items in the set is considerably large, then we can form a frequency distribution and determine the mode.

Example 1 :

Calculate the mode from the data given below :

2, 7, 11, 1, 13, 12, 17, 12, 19, 12, 39, 14, 12, 11, 10

The numbers given above are arranged in ascending order as follows :

Space for hints

1, 2, 7, 10, 11, 11, 12, 12, 12, 12, 13, 14, 17, 19, 39

The value 12 occurs four times; 11 occurs two times; and all the other values occur once. Thus 12 occurs greater number of times than others and it is the modal value.

Answer : Mode = 12

Example 2 :

Consider the following array and determine the mode.

7, 10, 10, 11, 12, 14, 14, 17, 19, 21.

Values 10 and 14 occur twice and all the other values occur once. Thus 10 and 14 occur greatest number of times than others. Hence, 10 as well as 14 can be taken as the modal values.

Answer : The modal values are 10 and 14. The given set is 'bi-modal'.

Example 3 :

Calculate the mode of the following :

2, 5, 5, 7, 11, 11, 11, 11, 14, 15, 15, 15, 15, 20, 23, 23, 23, 25, 25, 25, 25.

Value 11 occurs four times

Value 15 occurs four times

Value 25 occurs four times

Value 23 occurs three times

Value 5 occurs two times

The rest of the values occur once.

All the three values viz., 11, 15 and 25 occur greater number of times than others. Hence, all the three values viz., 11, 15 and 25 can be taken as the mode. Thus the given set has three modal values and is 'tri-modal'.

Answer : The modal values are 11, 15 and 25, The given set is 'tri-modal'.

5.3 Calculation of Mode for Discrete Frequency Distribution :

In a discrete frequency distribution, the value of the items having greatest frequency is found out. It is the value of the mode of the given distribution. Consider the following discrete frequency table.

Size of item	Frequency
10	3
12	4
14	12
16	5
18	2
20	1

The greatest frequency is 12. The value of the item having this greatest frequency is 14.

\therefore Mode = 14.

5.4 Calculation of Mode for Continuous Frequency Distribution :

(A) Crude method :

The maximum frequency of the given distribution is found out first. The class having the maximum frequency is found out next. This class is called the 'modal class'. The midvalue of the modal class is found out. The midvalue is taken as the value of the mode.

In this method the implicit assumption we make is that the greatest concentration of frequency of modal class is at the midpoint of the modal class. But usually this assumption is not true and hence the value of the mode obtained by this method is not accurate. Also the above method is adopted only when the given distribution is clearly uni-modal.

Example 4 :

Find the mode for the following data.

Weight (in lbs)	Frequency
90 - 100	10
100 - 110	37
110 - 120	65
120 - 130	80
130 - 140	51
140 - 150	35
150 - 160	18
160 - 170	4

The maximum frequency = 80

The class having maximum frequency is '120-130'

∴ The modal class is '120-130'

The midvalue of the modal class = $\frac{120+130}{2} = \frac{250}{2} = 125$

Hence, Mode = 125 Lbs.

Example 5 :

Calculate the mode of the following data.

Life (hours)	Frequency of lamps
0 - 400	3
400 - 800	12
800 - 1200	40
1200 - 1600	41
1600 - 2000	27
2000 - 2400	13
2400 - 2800	9
2800 - 3200	4

Maximum frequency = 41

The class having the maximum frequency is '1200 - 1600'

$$\text{The midvalue of the modal class} = \frac{1200 + 1600}{2} = \frac{2800}{2} = 1400$$

Hence, Mode = 1400 hours

Answer : Modal life of the lamps = 1400 hours.

(B) Mode calculated by giving weightage to the neighbouring frequencies of the modal class :

The value of the mode is considerably affected by the frequencies of the classes neighbouring to the modal class. If the frequencies of the class preceding the modal class is greater than the frequency of the class succeeding the modal class then the modal value will be nearer to the lower limit of the modal class. On the other hand, if the frequency of the class preceding the modal class is less than the frequency of the class succeeding the modal class then the modal value will be nearer to the upper limit of the modal class. Therefore, while giving the formula to calculate the mode, the frequencies of the classes preceding and succeeding the modal class are taken into account.

Consider the frequency distribution given below :

Weight (in lbs)	Frequency
90 - 100	10
100 - 110	37
110 - 120	65
120 - 130	80
130 - 140	51
140 - 150	35
150 - 160	18
160 - 170	4

In this case, the modal class is '120–130' because it has the highest frequency viz., 80.

The true lower limit of the modal class is denoted by 'l'

Here, $l = 120$

The class preceding the modal class is '110 – 120'

The frequency of the class '110–120' is 65

The frequency viz., the frequency of the class preceding the modal class is denoted by ' f_1 '.

$$\therefore f_1 = 65;$$

The class succeeding the modal class is '130–140'.

The frequency of the class '130–140' is 51.

The frequency of the class succeeding the modal class is denoted by ' f_2 '.

$$\therefore f_2 = 51.$$

The width of the modal class = $130 - 120 = 10$

The width of the modal class is denoted by ' c '

$$\therefore c = 10$$

Now, we give the formula to calculate the mode as follows

$$\text{Mode} = l + \frac{cf_2}{f_1 + f_2}$$

Where l = true lower limit of the modal class,

c = width of the modal class

f_1 = frequency of the class preceding the modal class.

f_2 = frequency of the class succeeding the modal class.

In our example

$$l = 120, \quad c = 10, \quad f_1 = 65, \quad f_2 = 51$$

$$\therefore \text{Mode} = 120 + \frac{10 \times 51}{65 + 51}$$

$$= 120 + \frac{510}{116}$$

$$= 120 + \frac{255}{58} = 124.4 \text{ lbs.}$$

Example 6 :

Consider the table given under Example 6 and calculate the mode using the formula given above.

Life (hours)	Frequency of lamps
0 - 400	3
400 - 800	12
800 - 1200	40
1200 - 1600	41
1600 - 2000	27
2000 - 2400	13
2400 - 2800	9
2800 - 3200	4

Maximum frequency = 41

The class having the maximum frequency is '1200 – 1600'.

∴ The modal class is '1200 – 1600'.

l = true lower limit of the modal class = 1200

c = width of the modal class

= 1600 – 1200 = 400

f_1 = frequency of the class preceding the modal class

= frequency of the class '800–1200' = 40

f_2 = frequency of the class succeeding the modal class

= frequency of the class '1600–2000' = 27

$$\therefore \text{Mode} = l + \frac{cf_2}{f_1 + f_2}$$

$$= 1200 + \frac{400 \times 27}{40 + 27} = 1200 + \frac{10800}{67}$$

$$= 1200 + 161.2 \text{ (approx.)} = 1361.2 \text{ hours}$$

Answer :

$$\text{Mode} = 1361.2 \text{ hours.}$$

Example 7 :

Calculate the mode of the following frequency distribution.

Space for hints

Class	Frequency
40 - 60	9
60 - 80	11
80 - 100	14
100 - 120	20
120 - 140	15
140 - 160	10

Highest frequency = 20

The class having the highest frequency is '100 - 120'

Modal class is '100 - 120'

l = True lower limit of the modal class = 100

c = width of the modal class = $120 - 100 = 20$

f_1 = frequency of the class preceding the modal class
= frequency of the class '80 - 100' = 14

f_2 = frequency of the class succeeding the modal class
= frequency of the class '120 - 140' = 15

$$\text{Mode} = l + \frac{cf_2}{f_1 + f_2}$$

$$= 100 + \frac{20 \times 15}{14 + 15}$$

$$= 100 + \frac{300}{29}$$

$$= 100 + 10.34 \text{ (approx.)} = 110.34$$

Answer : **Mode = 110.34**

(C) Mode calculated by taking the differences of neighbouring frequencies into consideration

The formula to calculate the mode can be given in terms of (i) the differences between the highest frequency and frequency of the class preceding the modal class and (ii) the difference between the highest frequency and the frequency of the class succeeding the modal class. Consider the frequency distribution given under.

Example 8 :

Weight (in lbs)	Frequency
90 - 100	10
100 - 110	37
110 - 120	65
120 - 130	80
130 - 140	51
140 - 150	35
150 - 160	18
160 - 170	47

In this distribution the modal class is '120 - 130'

Here also the true lower limit of the modal class is denoted by l

$$\therefore l = 120$$

The difference between the highest frequency and the frequency of the class preceding the modal class is denoted by ' d_1 '. In our example, highest frequency = 80.

Frequency of the class preceding the modal class = 65.

$$\therefore d_1 = 80 - 65 = 15$$

The difference between the highest frequency and the frequency of the class succeeding the modal class is denoted by ' d_2 '.

In our example, highest frequency = 80

Frequency of the class succeeding the modal class = 51

$$d_2 = 80 - 51 = 29$$

As before, the width of the modal class is denoted by 'c'.

$$c = 130 - 120 = 10$$

Now, the formula is given as follows :

$$\text{Mode} = l + \frac{cd_1}{d_1 + d_2}$$

Where l = true lower limit of the modal class

c = width of the modal class

d_1 = difference between highest frequency and the frequency of the class just above the modal class.

d_2 = difference between the highest frequency and the frequency of the class just below the modal class.

In our example, $l = 120$, $c = 10$, $d_1 = 15$, $d_2 = 29$

$$\text{Mode} = 120 + \frac{10 \times 15}{15 + 29}$$

$$= 120 + \frac{150}{44}$$

$$= 120 + 3.4 \text{ (approx.)} = 123.4 \text{ lbs.}$$

This formula viz., $\text{Mode} = l + \frac{cd_1}{d_1 + d_2}$ is better than the formula viz.,

$\text{Mode} = l + \frac{cf_2}{f_1 + f_2}$ because usually the former gives the value of the mode accurately.

Example 9 :

Consider the table given under Example 6, and calculate the mode using the formula containing the differences of frequencies.

Maximum frequency = 41.

The class having this maximum frequency is '1200 - 1600'.

Modal class is '1200 - 1600'

l = lower limit of the modal class = 1200

Space for hints

Check your Progress

9. Give any one formula to calculate mode of a continuous frequency distribution.

$$c = \text{width of the modal class} = 1600 - 1200 = 400$$

d_1 = difference between the highest frequency and the frequency of the class just above the modal class.

$$= \text{difference between 41 and the frequency of the class '800-1200'} = 41 - 40 = 1$$

d_2 = difference between the highest frequency and the frequency of the class just below the modal class

$$= \text{difference between 41 and the frequency of the class '1600-2000'} = 41 - 27 = 14.$$

$$\text{Mode} = l + \frac{cd_1}{d_1 + d_2}$$

$$= 1200 + \frac{400 \times 1}{1 + 14}$$

$$= 1200 + \frac{400}{15}$$

$$= 1200 + 26.67 \text{ (approx.)}$$

$$= 1226.67 \text{ hours.}$$

Answer : **Mode = 1226.67 hours**

Example 10 :

Consider the table given under example 8 and calculate the mode using the formula containing the differences of frequencies.

Highest frequency = 20

Modal class is '100 - 200'

l = true lower limit of the modal class = 100

c = width of the modal class = 120 - 100 = 20

The frequency of the class just above the modal class = 14

$$d_1 = 20 - 14 = 6$$

Frequency of the class just below the modal class = 15

$$d_2 = 20 - 15 = 5.$$

$$\begin{aligned}\text{Mode} &= 100 + \frac{cd_1}{d_1 + d_2} \\ &= 100 + \frac{20 \times 6}{6 + 5} \\ &= 100 + \frac{120}{11} \\ &= 100 + 10.9 \text{ (approx.)} \\ &= 110.9\end{aligned}$$

Answer : **Mode = 110.9**

In all the examples we have considered above, the modal class is somewhere in the middle of the distribution.

Therefore, we have classes preceding the succeeding the modal class. But this need not be the case always. The highest frequency may occur in the first class itself. Therefore the first class is the modal class and there is no class preceding the modal class.

∴ The values of f_1 and d_1 do not exists.

The two formulae we have given above cannot be used.

In such a case, the midvalue of the first class is taken as the mode. Consider the following distribution:

Class	Frequency
20 - 40	20
40 - 60	14
60 - 80	11
80 - 100	9
100 - 120	8

The highest frequency = 20

The class having the highest frequency is the first class viz., '20 - 40'

∴ The midvalue of the first class is taken as the mode.

$$\text{Midvalue of the first class} = \frac{20+40}{2} = \frac{60}{2} = 30$$

$$\therefore \text{Mode} = 30$$

In certain cases, the highest frequency may occur in the last class. Therefore, the last class is the modal class and there is no class succeeding the modal class.

\therefore The values of f_2 and d_2 do not exist.

$$\therefore \text{The two formulae viz., } \text{Mode} = l + \frac{cf_2}{f_1 + f_2} \text{ and } \text{Mode} = l + \frac{cd_1}{d_1 + d_2}$$

cannot be used to calculate the mode.

In such a case, the midvalue of the last class is taken as the mode.

Consider the following distribution :

Class	Frequency
20 - 40	8
40 - 60	11
60 - 80	9
80 - 100	14
100 - 120	20

The maximum frequency = 20

The class having maximum frequency is the last class viz., '100 - 120'

\therefore The midvalue of the last class viz., 110 is taken as the mode.

$$\text{Mode} = 110$$

Thus, when the maximum frequency occurs in any of the extreme classes of the given distribution viz., either in the first class or in the last class mode can be calculated only by the crude method.

Sometimes, the given distribution may have open intervals at its end and the maximum frequency may occur in any one of the end classes. In such a case, mode cannot be found out even crudely because the midvalue of an open class is indeterminate.

Sometimes, the given distribution may have open intervals at its ends but the highest frequency may occur somewhere in the middle of the distribution. In such a case, we can apply either of the two formulae,

$$\text{Mode} = l + \frac{cf_2}{f_1 + f_2}$$

$$\text{Mode} = l + \frac{cd_1}{d_1 + d_2}$$

and calculate the mode. In this case, the open intervals do not pose any difficulty in calculating the mode.

Consider the following example.

Example 11 :

Calculate the mode of the following distribution :

Class	Frequency
Below 50	5
50 - 70	10
70 - 90	19
90 - 110	20
Above 130	16

In this distribution the first class as well as the last class are open classes. But the highest frequency does not belong to any of these two classes.

The highest frequency = 30 and the class containing the highest frequency is '90 - 110'.

∴ The modal class is '90 - 110'

l = lower limit of the modal class = 90

c = width of the modal class = $110 - 90 = 20$

f_1 = frequency of the class just above the modal class = 19

f_2 = frequency of the class just below the modal class = 20

$$\text{Mode} = l + \frac{cf_2}{f_1 + f_2}$$

$$= 90 + \frac{20 \times 20}{19 + 20} = 90 + \frac{400}{39}$$

$$= 90 + 10.26 \text{ (approx.)} = 100.26$$

Mode of the given distribution can also be calculated using the other formula as shown below.

$$d_1 = 30 - 19 = 11$$

$$d_2 = 30 - 20 = 10$$

$$\text{Mode} = l + \frac{cd_1}{d_1 + d_2}$$

$$= 90 + \frac{20 \times 11}{11 + 10} = 90 + \frac{220}{21}$$

$$= 90 + 10.5 \text{ (approx.)} = 100.5$$

In certain cases, the maximum frequency may occur in two or more consecutive classes. Such distribution is considered to be uni-modal. The midvalue of the entire range covered by these consecutive classes is found out. This midvalue is taken as the mode.

Consider the following frequency distribution :

Class	Frequency
0 - 10	4
10 - 20	9
20 - 30	15
30 - 40	15
40 - 50	15
50 - 60	8
60 - 70	1

The maximum frequency = 15

It occurs in the three consecutive classes viz., '20-30', '30-40' and '40-50'

The range covered by these three classes is 20 to 50

$$\text{The midvalue of this range} = \frac{50 + 20}{2} = \frac{70}{2} = 35$$

\therefore Mode = 35

If the maximum frequency occurs in two or more classes which are not consecutive, then we have to calculate the different modes separately.

Space for hints

Consider the following example :

Class	Frequency
10 - 12	10
12 - 14	40
14 - 16	20
16 - 18	10
18 - 20	40
20 - 22	16
22 - 24	6

Maximum frequency = 40

It occurs in the two classes '12-14' and '18-20'

These two classes are not consecutive.

∴ Separate modes are to be calculated.

Mode corresponding to class '12-14'

l = true lower limit of the modal class = 12

c = width of the modal class = $14 - 12 = 2$

f_1 = frequency of the class just above the modal class = 10

f_2 = frequency of the class just below the modal class = 20

$$\text{Mode} = l + \frac{cf_2}{f_1 + f_2}$$

$$= 12 + \frac{2 \times 20}{10 + 20}$$

$$= 12 + \frac{40}{30}$$

$$= 12 + 1.33 = 13.33$$

Space for hints

(ii) Mode corresponding to the class '18-20'

$$l = 18 \quad c = 20 - 18 = 2 \quad f_1 = 10 \quad f_2 = 16$$

$$\begin{aligned} \text{Mode} &= l + \frac{cf_2}{f_1 + f_2} \\ &= 18 + \frac{2 \times 16}{10 + 16} \\ &= 18 + \frac{32}{26} \\ &= 18 + 1.23 = 19.23 \end{aligned}$$

∴ The two modes are 13.33 and 19.23

If in the given distribution the class intervals are not true class intervals, they must be converted into true class intervals first. Then only we should apply the formula to calculate the mode.

Consider the following example.

Example 12 :

Calculate the mode

Class	Frequency
10 - 19	8
20 - 29	10
30 - 39	16
40 - 49	20
50 - 59	14
60 - 69	6

The class intervals of the distribution given above are not true class intervals.

∴ First of all they are converted into true class intervals and the table is given as follows :

Class	Frequency
9.5 - 19.5	8
19.5 - 29.5	10
29.5 - 39.5	16
39.5 - 49.5	20
49.5 - 59.5	14
59.5 - 69.5	6

Space for hints

Maximum frequency = 20

The class having the maximum frequency is '39.5 – 49.5'

Modal class is '39.5 – 49.5'

$$\therefore l = 39.5 \quad c = 49.5 - 39.5 = 10 \quad f_1 = 16 \quad f_2 = 14$$

$$\text{Mode} = l + \frac{cf_2}{f_1 + f_2}$$

$$= 39.5 + \frac{10 \times 14}{16 + 14} = 39.5 + \frac{140}{30}$$

$$= 39.5 + 4.66 = 44.16$$

D) Calculation of Mode using Mean and Median :

If the mean and median of a given distribution are known, then the mode can be calculated using the following relation viz.,

$$(\text{Mean} - \text{Mode}) = 3(\text{Mean} - \text{Median})$$

Example 13 :

The mean and median of a distribution are 125.73 and 124.75 respectively. Calculate the mode.

$$(\text{Mean} - \text{Mode}) = 3(\text{Mean} - \text{Median})$$

$$(125.73 - \text{Mode}) = 3(125.73 - 124.75)$$

$$(125.73 - \text{Mode}) = 3 \times .98$$

$$(125.73 - \text{Mode}) = 2.94$$

$$\therefore \text{Mode} = 125.73 - 2.94 = 122.79$$

Check your Progress

10. . State the relationship existing among Median and Mode.

6. GEOMETRIC MEAN

6.1 Definition :

Geometric Mean is defined as the n^{th} root of the product of the given items where n stands for the total number of items given.

6.2 Calculation of geometric mean from ungrouped data :

Let us assume that n values of a variable are given. To get the geometric mean of this set of n values, the product of these n values is obtained first. Then the n^{th} root of the product is found out. The value of the n^{th} root of the product gives the value of the geometric mean of the given set of n values.

Suppose, the given variable is 'age'. The ages of two persons are given as 16 and 25. Therefore, 16 and 25 are two values of the given variable 'age'. To get the geometric mean of these two values we calculate their product first.

$$16 \times 25 = 400$$

Since two values viz., 16 and 25 are given, $n = 2$

The value of the geometric mean is n^{th} root of the product of the given values. Since $n = 2$, in our example, we find the square root of the product of the given values viz., 400.

$$\sqrt{400} = \sqrt{20 \times 20} = 20$$

Therefore, 20 is the geometric mean of the two values viz., 16 and 25.

Suppose, one more value say, 20 is given in addition to the two values 16 and 25 given in the previous example.

Now, we have three values of the given variable viz., 'age'. Hence $n = 3$.

To get the geometric mean, the three values are multiplied with each other and product is obtained as follows :

$$16 \times 25 \times 20 = 8000.$$

The n^{th} root of the product is found out next.

Since, $n = 3$ we find the cube root of the product viz., 8,000.

$$\sqrt[3]{8000} = \sqrt[3]{20 \times 20 \times 20} = 20$$

\therefore 20 is the geometric mean of the three values given viz., 16, 25 and 20.

Suppose, four values of the variable viz., 'age' are given. Let the four values be 9, 18, 16, 8.

n = number of values given = 4

**Check your
Progress**

11. Define G.M.

Geometric mean is the n^{th} root of the product of the n values given.

Space for hints

Since here $n = 4$, the product of the four values is found out and 4th root of the product is calculated.

Product of the four values = $9 \times 18 \times 16 \times 8$

4th root of this product

$$\begin{aligned}
 &= \sqrt[4]{9 \times 18 \times 16 \times 8} \\
 &= \sqrt[4]{(3 \times 3) \times (3 \times 3 \times 2) \times (2 \times 2 \times 2 \times 2) \times (2 \times 2 \times 2)} \\
 &= \sqrt[4]{(3 \times 3 \times 3 \times 3) \times (2 \times 2 \times 2 \times 2) \times (2 \times 2 \times 2 \times 2)} \\
 &= 3 \times 2 \times 2 = 12
 \end{aligned}$$

\therefore Geometric mean of the four values given = 12

Suppose a set of values are given.

Let the first value in the set be denoted by ' x_1 ', the second value be denoted by ' x_2 ' and so on.

The final value in the set is denoted by ' x_n '.

The number of values given is equal to n . Now, the geometric mean of the set of values is given by the following formula.

$$G = \sqrt[n]{x_1 x_2 \dots x_n}$$

Where G denotes the geometric mean

Consider the following example.

Example 1 :

Calculate the geometric mean of the following set of values 125, 64, 216

n = number of values given = 3

x_1 = first value in the given set = 125

x_2 = second value in the given set = 64

x_3 = third value which is also the final value in the given set = 216

$$\begin{aligned}
 \therefore G &= \sqrt[3]{125 \times 64 \times 216} \\
 &= \sqrt[3]{(5 \times 5 \times 5) \times (4 \times 4 \times 4) \times (6 \times 6 \times 6)}
 \end{aligned}$$

$$= \sqrt[3]{(5 \times 4 \times 6) \times (5 \times 4 \times 6) \times (5 \times 4 \times 6)}$$

$$= 5 \times 4 \times 6 = 120$$

Answer : Geometric mean = 120

6.3 Alternative Formula to Get Geometric Mean from ungrouped Data

In the examples we have given above we were able to find out the root value easily without the help of logarithmic tables. The reason is that the number of values given is small and their sizes are also small. But often this will not be the case. We have to use logarithmic tables in order to find out the root value of the product. Hence, usually geometric mean is given in logarithmic terms as follows :

$$\log G = \frac{\log x_1 + \log x_2 + \dots + \log x_n}{n}$$

Where G denotes geometric mean,

x_1, x_2, \dots, x_n denote the given values of a variable

n denotes the number of values given.

The sum $(\log x_1 + \log x_2 + \dots + \log x_n)$ is denoted by $\Sigma \log x$ where 'Σ' stands for 'summation of'.

$$\therefore \log G = \frac{\Sigma \log x}{n}$$

$$G = \text{Anti-log} \left\{ \frac{\Sigma \log x}{n} \right\}$$

The process of calculation of geometric mean can be given in a summarized form as follows :

1. Log of each value given is found out using log tables.
2. All the log values are summed up and the value of $\Sigma \log x$ is obtained.
3. The sum $\Sigma \log x$ is divided by n where n denotes the number of values given.

Thus the value of $\frac{\Sigma \log x}{n}$ is obtained.

4. Anti-log value of $\frac{\Sigma \log x}{n}$ is found out using anti-log tables.

This gives the value of the geometric mean.

Check your Progress

12. Give the formula to calculate G.M. from ungrouped data.

For example, consider the following numbers :

Space for hints

331, 411, 251, 713, 812

First, log of each value given is found out using log-table as follows :

$$\log 331 = 2.5198$$

$$\log 411 = 2.6138$$

$$\log 251 = 2.3997$$

$$\log 713 = 2.8531$$

$$\log 812 = 2.9096$$

All the log values are added to give the value of $\Sigma \log x$

$$\begin{aligned}\therefore \Sigma \log x &= 2.5198 + 2.6138 + 2.3997 + 2.8531 + 2.9096 \\ &= 13.2960\end{aligned}$$

The value of $\Sigma \log x$ is divided by n where n denotes the number of values given.

In our example, five values are given

$$\therefore n = 5$$

\therefore Value of $\Sigma \log x$ viz., 13.2960 is divided by 5.

$$\frac{\Sigma \log x}{n} = \frac{13.2960}{5} = 2.6592$$

Anti-log value of $\frac{\Sigma \log x}{n}$ is found out

(i.e.) anti-log value of 2.6592 is found out from anti-log tables.

$$\text{Anti-log}(2.6592) = 456.2$$

$$\therefore G = 456.2$$

(i.e.,) Geometric mean = 456.2

Example 2 :

Calculate the geometric mean of the set of values given below :

28, 54.6, .987, 3.493, .0132, .05408

Using log tables, log value of each number given above is found out.

$$\log 28 = 1.4472$$

$$\log 54.6 = 1.7372$$

$$\log .987 = \bar{1}.9943$$

$$\log 3.493 = 0.5432$$

$$\log .0132 = \bar{2}.1206$$

$$\log .05408 = \bar{2}.7330$$

All the log values are added to give the value of $\Sigma \log x$

$$\begin{aligned}\Sigma \log x &= 1 + .4472 + 1.7372 + \bar{1}.9943 + .5432 + \bar{2}.1206 + \bar{2}.7330 \\ &= 1 + .4472 + 1 + .7372 - 1 + .9943 + .5432 - 2 + .1206 - 2 + .7330 \\ &= 1+1-1-2-2+.4472+.7372+.9943+.5432+.1206+.7330 \\ &= 2-5+3.5755 \\ &= -3+3.5755 \\ &= .5755\end{aligned}$$

Now $\Sigma \log x$ is divided by n where n denotes the number of values given.

In our problem we are given 6 values.

$$\therefore n = 6$$

The value of $\Sigma \log x$ viz., .5755 is divided by 6 to get the value of $\frac{\Sigma \log x}{n}$

$$\frac{\Sigma \log x}{n} = \frac{.5755}{6} = .0959$$

Using anti-log tables, anti-log value of $\frac{\Sigma \log x}{n}$ viz., .0959 is found out.

$$\text{Anti-log} (.0959) = 1.248$$

$$G = 1.248$$

Answer : Geometric mean = 1.248

6.4 Calculation of geometric mean from discrete frequency distribution :

Space for hints

Consider the following discrete frequency distribution.

Value of item	Frequency
2	4
5	9
6	11
8	6

Value of the first item viz., 2 is denoted by x_1 .

Value of the second item viz., 5 is denoted by x_2 .

Value of third item viz., 6 is denoted by x_3 .

Value of fourth item viz., 8 is denoted by x_4 .

The frequency of the first item viz., 4 is denoted by f_1 , frequency of the second item viz., 9 is denoted by f_2 , of the third item viz., 11 by f_3 and of the fourth item viz., 6 by f_4 .

Now, the geometric mean is given as follows.

$$G = \sqrt[N]{x_1^{f_1} \cdot x_2^{f_2} \cdot x_3^{f_3} \cdot x_4^{f_4}}$$

Where G denotes the geometric mean and N denotes the sum of frequencies viz., $(f_1 + f_2 + f_3 + f_4)$

In general, the formula to calculate the geometric mean of discrete frequency distribution is given as follows :

$$G = \sqrt[N]{x_1^{f_1} \cdot x_2^{f_2} \cdot \dots \cdot x_n^{f_n}}$$

Where G denotes geometric mean

x_1 denotes the value of the first item

x_2 denotes the value of the second item

.....

.....

x_n denotes the value of the final item

f_1 denotes the frequency of the first item

f_2 denotes the frequency of the second item

.....

.....

f_n denotes the frequency of the final item

N denotes the total sum of frequencies viz., $(f_1 + f_2 + f_3 + \dots + f_n)$.

6.5 Alternative Formula to Get Geometric Mean from Discrete Frequency Distribution :

In the case of discrete frequency distribution, it will be tedious to calculate the geometric mean using the formula given above. With the help of logarithmic table we can easily calculate the value of the geometric mean.

Therefore, the formula to calculate the geometric mean is given in terms of logarithms below. The formula given above is seldom used to calculate geometric mean. Only the formula given below in logarithmic terms is generally used. Hence, the students are advised to use the formula given below whenever they have to calculate the geometric mean.

$$\log G = \frac{f_1 \log x_1 + f_2 \log x_2 + \dots + f_n \log x_n}{N}$$

$$\text{Where } N = f_1 + f_2 + f_3 + \dots + f_n.$$

Usually, the sum $(f_1 \log x_1 + f_2 \log x_2 + \dots + f_n \log x_n)$ is denoted by $\Sigma f \log x$ and the sum of frequencies viz., $(f_1 + f_2 + \dots + f_n)$ is denoted by Σf . So, $N = \Sigma f$

$$\therefore \log G = \frac{\Sigma f \log x}{N}$$

$$G = \text{Anti-log} \left\{ \frac{\Sigma f \log x}{N} \right\}$$

Method of calculation of geometric mean of a discrete frequency distribution using the formula given above viz.,

$$G = \text{Anti-log} \left\{ \frac{\Sigma f \log x}{N} \right\}$$

can be given in a summarized form as follows :

1.

Using the tables, log of each value of the item given in the table is found out.
2.

Log value of each item is multiplied by the corresponding frequency. If x is the value of an item its log value is $\log x$. If f is the frequency of the item x , then f and $\log x$ are multiplied and $f \log x$ value is obtained. These products are summed up to give $\Sigma f \log x$.
3.

All the frequencies are added to give the value of $\Sigma f = N$
4.

Value of $\Sigma f \log x$ is divided by N and the value $\frac{\Sigma f \log x}{N}$ is obtained.
5.

Using anti-log tables, anti-log of $\frac{\Sigma f \log x}{N}$ is found out. This gives the value of geometric mean.

Consider the table 1, given above. For each value of item log value is found out.

$$\log 2 = 0.3010$$
$$\log 5 = 0.6990$$
$$\log 6 = 0.7782$$
$$\log 8 = 0.9031$$

Log, value of each item is multiplied by the corresponding frequency and therefore, we get

$$4 \times \log 2 = 4 \times .3010 = 1.2040$$
$$9 \times \log 5 = 9 \times .6990 = 6.2910$$
$$11 \times \log 6 = 11 \times .7782 = 8.5602$$
$$6 \times \log 8 = 6 \times .9031 = 5.4186$$

Log. values and the products of log. values and their corresponding frequencies can be given in a tabular form as follows.

Value of item x	Frequency f	$\log x$	$f \log x$
2	4	.3010	1.2040
5	9	.6990	6.2910
6	11	.7782	8.5602
8	6	.9031	5.4186
Total	30		21.4738

Space for hints

$$N = \Sigma f = 30$$

$$\Sigma f \log x = 21.4738$$

$$\frac{\Sigma f \log x}{N} = \frac{21.4738}{30} = .71579 = .7158 \text{ (Approx.)}$$

Anti-log $\left(\frac{\Sigma f \log x}{N} \right)$ gives the value of geometric mean

(i.e.) Anti-log (0.7158) gives the value of geometric mean.

$$\text{Anti-log}(0.7158) = 5.198$$

$$\text{Geometric mean} = 5.198$$

Example 3 :

Calculate the geometric mean

Value of the item	Frequency
.75	5
1.3	8
1.8	10
2.0	14
2.2	6

For each value of item log value is found out

$$\log .75 = \bar{1}.8751$$

$$\log 1.3 = .1139$$

$$\log 1.8 = .2553$$

$$\log 2 = .3010$$

$$\log 2.2 = .3424$$

Log. value of each item multiplied by the corresponding frequency as follows

$$5 \times \log .75 = 5 \times (\bar{1}.8751)$$

$$= 5 \times (-1 + .8751)$$

$$= -5 + 4.3755$$

$$= -1 + .3755$$

$$= \bar{1}.3755$$

$$8 \times \log 1.3 = 8 \times .1139 = .9112$$

$$10 \times \log 1.8 = 10 \times .2553 = 2.5530$$

$$14 \times \log 2 = 14 \times .3010 = 4.2140$$

$$6 \times \log 2.2 = 6 \times .3424 = 2.0544$$

Log. values and the products of log values and the corresponding frequencies can be given in a tabular form as follows :

Value of item x	Frequency f	log x	f logx
.75	5	̄1.8751	̄1.3755
1.3	8	.1139	.9112
1.8	10	.2553	2.5530
2.0	14	.3010	4.2140
2.2	6	.3424	2.0544
Total	43		9.1081

$$N = \Sigma f = 43 \quad \Sigma f \log x = 9.1081$$

$$\frac{\Sigma f \log x}{N} = \frac{9.1081}{43} = .2118$$

Anti-log $\left(\frac{\Sigma f \log x}{N} \right)$ gives the value of geometric mean (i.e.) Anti-log (0.2118) gives the value of geometric mean.

$$\text{Anti-log } .2118 = 1.629$$

$$\therefore \text{Geometric mean} = 1.629.$$

6.6 Calculation of Geometric Mean from continuous frequency distribution :

Consider the following continuous frequency distribution.

Class interval	Frequency
0 - 10	5
10 - 20	15
20 - 30	18
30 - 40	13
40 - 50	4

To calculate the geometric mean, midvalue of each class is found out first.

The midvalue of the first class is denoted by x_1 . In our example, midvalue of the first class is 5 and is denoted by x_1 .

The midvalue of the second class is denoted by x_2 . In our example, midvalue of the second class is 15 and it is denoted by x_2 and so on.

The midvalue of the final class is denoted by x_n . In our example, the final class is 40-50 and its midvalue viz., 45 is denoted by x_n .

Frequency of the first class is denoted by f_1 . In our example frequency 5 is denoted by f_1 .

Frequency of the second class is denoted by f_2 and so on. In our example frequency 15 is denoted by f_2 and so on.

Frequency of the final class is denoted by f_n . In our example frequency of the final class viz., 4 is denoted by f_n .

Now the formula to calculate the geometric mean is given as follows :

$$G = \sqrt[N]{x_1^{f_1} \cdot x_2^{f_2} \cdot \dots \cdot x_n^{f_n}}$$

where G denotes geometric mean,

N denotes the sum of frequencies viz., $(f_1 + f_2 + \dots + f_n)$.

6.7 Alternative Formula to Get Geometric Mean from continuous frequency Distribution :

As we have stated in the previous case here also the formula given above is seldom used to calculate the geometric mean. We have given below the formula to calculate the geometric mean in logarithmic terms and mostly this formula is used to calculate the geometric mean. Therefore, students are advised to use the formula given below whenever they have to calculate the geometric mean of a continuous frequency distribution.

$$\log G = \frac{f_1 \log x_1 + f_2 \log x_2 + \dots + f_n \log x_n}{N}$$

where $N = f_1 + f_2 + f_3 + \dots + f_n$.

Usually, the sum $(f_1 \log x_1 + f_2 \log x_2 + \dots + f_n \log x_n)$ is denoted by $\Sigma f \log x$ and the sum of frequencies viz., $(f_1 + f_2 + \dots + f_n)$ is denoted by Σf . So, $N = \Sigma f$

$$\therefore \log G = \frac{\sum f \log x}{N}$$

$$G = \text{Anti-log} \left\{ \frac{\sum f \log x}{N} \right\}$$

Method of calculation of geometric mean of a discrete frequency distribution can be given in a summarized form as follows :

1. Midvalue of each class is found out first.
2. Logarithm of each midvalue is found out next. If x is the midvalue of a class, value of $\log x$ is found out.
3. Logarithm of each midvalue is multiplied by the corresponding frequency. If f is the frequency corresponding to the midvalue x , then $\log x$ is multiplied by f . The product is $f \log x$.
4. All the values of $f \log x$ are summed up and the sum is denoted by $\sum f \log x$.
5. All the frequencies are summed up and the sum is denoted by N .
6. $\sum f \log x$ is divided by N and the value of $\frac{\sum f \log x}{N}$ is found out.
7. Anti-log value of $\frac{\sum f \log x}{N}$ is found out. This gives the value of geometric mean.

Consider the table 2 given above. The midvalues of the class intervals given are 5, 15, 25, 35 and 45 respectively. Logarithm of each midvalue is found out as follows.

$$\log 5 = 0.6990$$

$$\log 15 = 1.1761$$

$$\log 25 = 1.3979$$

$$\log 35 = 1.5441$$

$$\log 45 = 1.6532$$

Log. of each midvalue is multiplied by the corresponding frequency as follows :

$$5 \times \log 5 = 5 \times .6990 = 3.4950$$

$$15 \times \log 15 = 15 \times 1.1761 = 17.6415$$

$$18 \times \log 25 = 18 \times 1.3979 = 25.1622$$

Space for hints

$$13 \times \log 35 = 13 \times 1.5441 = 20.0733$$

$$4 \times \log 45 = 4 \times 1.6532 = 6.6128$$

Log. values and the products of log values and their corresponding frequencies can be given in a tabular form as follows.

Mid value x	Frequency f	log x	f log x
5	5	.6990	3.4950
15	15	1.1761	17.6415
25	18	1.3979	25.1622
35	13	1.5441	20.0733
45	4	1.6532	6.6128
Total	55		72.9848

$$N = \Sigma f = 55$$

$$\Sigma f \log x = 72.9848$$

$$\frac{\Sigma f \log x}{N} = \frac{72.9848}{55} = 1.3270 \text{ (Approx.)}$$

$$\text{Anti} - \log \left(\frac{\Sigma f \log x}{N} \right) = \text{Anti} - \log [1.3270] = 21.23$$

$$\therefore \text{Geometric mean} = 21.23$$

Example 4 :

Calculate the geometric mean.

Class interval	Frequency
0 - 4	4
4 - 8	10
8 - 12	12
12 - 16	6

Midvalue of each class is found out and given under x below.

Logarithm of each mid-value is found out and given under log x below.

Each value of log x is multiplied by the corresponding frequency and given under f log x below.

Mid value x	Frequency f	log x	f log x
2	4	0.3010	1.2040
6	10	0.7782	7.7820
10	12	1.0000	12.0000
14	6	1.1461	6.8766
Total	32		27.8626

$$N = \Sigma f = 32 \quad \Sigma f \log x = 27.8626$$

$$\frac{\Sigma f \log x}{N} = \frac{27.8626}{32} = .8707$$

$$\text{Anti-log} \left(\frac{\Sigma f \log x}{N} \right) = \text{Anti-log} [.8707]$$

$$= 7.425$$

$$\therefore \text{Geometric mean} = 7.425$$

6.8 Uses of Geometric Mean :

Geometric mean is the only useful average that can be employed to indicate rate of change. When percentage increases over a period of time are given, to find out the average percentage increase we must use only geometric mean.

Also when ratios of prices of two commodities over a period of time or some other ratios are given, to find out the average of the ratios given we must use only geometric mean.

6.9 Procedure to find out the average percentage increase when percentage increase in each period are given :

1. We assume the value at the beginning of each period to be 100
2. We add the percentage increase in each period to 100 and get the value at the end of each period. For example, if the increase in the first period is 10, then the value at the end of the first period is got by adding 10 to 100 and it is equal to 110.

If 5% is the increase in the second period, then, 105(=100 + 5) is the value at the end of the second period.

3. For the values at the end of each period, we find the geometric mean using the formula

$$G = \text{Anti-log} \left(\frac{\sum f \log x}{N} \right)$$

where G denotes geometric mean

x denotes the value at the end of a period

$\sum \log x$ denotes the sum of log of all x values.

N denotes the total number of periods given.

4. Value of G gives the average value at the end of each period. We subtract 100 from the value of G. The balance we get is the required average percentage or the average rate of increase.

$$\therefore \text{Average rate of increase} = G - 100$$

Example :

The price of a commodity increased by 5% from 1948 to 1949, 8% from 1949 to 1950 and 77% from 1950 to 1951. What is the average increase from 1948 to 1951?

Since percentage increases and not absolute increases are given in this problem only geometric mean should be used to find out the average increases.

We calculate the geometric mean as follows : We suppose that the price of the commodity at the beginning of each year to be equal to 100.

Assuming the price at the beginning of 1948 as 100, we get the price at the end of 1948 (ie., at the beginning of 1949) as $100 + 5 = 105$.

Assuming the price at the beginning of 1949 as 100, we get the price at the end of 1950 (ie., at the beginning of 1950), as $100 + 8 = 108$.

Assuming the price at the beginning of 1950 as 100, we get the price at the end of 1950 as $100 + 77 = 177$.

To find out the geometric mean of the prices at the end of each year, we use the formula

$$G = \text{Anti-log} \left(\frac{\sum \log x}{n} \right)$$

$$\log 105 = 2.0212$$

$$\log 108 = 2.0334$$

$$\log 177 = 2.2480$$

$$\text{Sum of the log values} = 2.0212 + 2.0334 + 2.2480 = 6.3026$$

$$\therefore \sum \log x = 6.3026$$

$$n = \text{number of values given} = 3$$

$$\frac{\sum \log x}{n} = \frac{6.3026}{3} = 2.1009$$

$$\begin{aligned} G &= \text{Anti-log} \left(\frac{\sum \log x}{n} \right) \\ &= \text{Anti-log} (2.1009) = 126.2 \end{aligned}$$

$$\therefore \text{Average of the prices at the end of each year} = 126.2$$

We have assumed the price at the beginning of each year to be 100.

We have got the average of the price at the end of each year to be 126.2

$$\text{Average increase in price} = 126.2 - 100 = 26.2\%$$

Example :

The population of a country increased by 20% in the first decade, 30% in the second decade and 45% in the third decade. What is the average rate of increase per decade in the population?

As we have assumed in the previous example, here also we assume the population at the beginning of each decade to be 100.

$$\therefore \text{Population at the end of the first decade} = 100 + 20 = 120$$

$$\text{Population at the end of the second decade} = 100 + 30 = 130$$

$$\text{Population at the end of the third decade} = 100 + 45 = 145$$

Geometric mean of the population at the end of the three decades given

above is obtained using the formula,

$$G = \text{Anti-log} \left(\frac{\sum \log x}{n} \right)$$

$$\log 120 = 2.0792$$

$$\log 130 = 2.1139$$

$$\log 145 = 2.1614$$

$$\sum \log x = 2.0792 + 2.1139 + 2.1614 = 6.3545$$

$$n = 3$$

$$\therefore \frac{\sum \log x}{n} = \frac{6.3545}{3}$$

$$= 2.1182$$

$$\text{Anti-log} \left(\frac{\sum \log x}{n} \right) = \text{Anti-log}(2.1182)$$

$$= 131.3$$

$$\therefore G = 131.3$$

(i.e.,) Average of populations at the end of each of the three decades = 131.3

Population at the beginning of each decade = 100.

Average rate of decade increase in the population

$$= 131.3 - 100 = 31.3\%$$

Example :

A manager having three shops found that the sales had increased by 125% in shop 'A', 75% in shop 'B' and 70% in shop 'C'. He concluded that average increase in sales was 90%. Is this correct? If not, how would you advise the manager to arrive at the correct average?

Manager's conclusion viz., the average increase in sales was 90% is wrong. When percentage increases are given, the average percentage increase will be given correctly only by geometric mean and not by arithmetic mean. But the manager has used arithmetic mean and got the average percentage increase as 90%. Therefore, the result is wrong.

The manager should find out the geometric mean of the percentage increase as we have done below.

Space for hints

Let us assume that before the sales have increased, the sales amount in each of the shops, A, B, C is equal to 100.

Now, after the sales have increased,

the sales amount in shop A = $100 + 125 = 225$

the sales amount in shop B = $100 + 75 = 175$

the sales amount in shop C = $100 + 70 = 170$

Geometric mean of these increased sales amounts is calculated using the formula

$$G = \text{Anti-log} \left(\frac{\sum \log x}{n} \right)$$

$$\log 225 = 2.3522$$

$$\log 175 = 2.2430$$

$$\log 170 = 2.2304$$

$$\therefore \sum \log x = 2.3522 + 2.2430 + 2.2304 = 6.8256$$

$$n = 3$$

$$\therefore \frac{\sum \log x}{n} = \frac{6.8256}{3} = 2.2752$$

$$\text{Anti-log} \left(\frac{\sum \log x}{n} \right) = \text{Anti-log} (2.2752) = 188.5$$

$$G = 188.5$$

(i.e.) Average increased sales = 188.5

The sales before the increase = 100

$$\therefore \text{Average increase in sales} = 188.5 - 100 = 88.5\%$$

The value of G gives the average value of each period with the assumption that value at the beginning of the period is 100. Therefore, if we subtract 100 from the value of G, we get the average percentage increase in value of the given variable.

Example :

From the following data find out the average rate of increase in population per decade.

Year	Population in millions
1881	250.2
1891	279.6
1901	283.9
1911	303.4

Here, absolute values are given, so we first find out the relative values as follows:

We express the value of population in 1881 to be equal to 100.

$$\therefore \text{Relative value of population in 1891} = \frac{279.6}{250.2} \times 100 = 111.8$$

We express the value of population in 1891 to be equal to 100; we find out the relative value of population in 1901 as

$$\frac{283.9}{279.6} \times 100 = 101.5$$

Similarly expressing the value in 1901 to be equal to 100 we get the relative value in 1911 as

$$\frac{303.4}{283.9} \times 100 = 106.9$$

For these three relative values viz., 111.8, 101.6 and 106.9 we find out the geometric mean using the formula.

$$G = \text{Anti-log} \left(\frac{\sum \log x}{n} \right)$$

$$\log 111.8 = 2.0483$$

$$\log 101.6 = 2.0068$$

$$\log 106.9 = 2.0288$$

$$\Sigma \log x = 2.0483 + 2.0068 + 2.0288 = 6.0839$$

$$n = 3$$

$$\frac{\Sigma \log x}{n} = \frac{6.0839}{3} = 2.0279$$

$$\text{Anti-log} \left(\frac{\Sigma \log x}{n} \right) = \text{Anti-log} (2.0279) = 106.6$$

$$\therefore G = 106.6$$

Now, we subtract 100 from the value of G and get $(106.6 - 100) = 6.6$

It is the average rate of increase in population.

Example :

The ratio of wheat to gram prices in 1960 is 2 and in 1965 is 4. What is the average ratio of wheat to gram prices?

When ratios are given and we have to find out the average ratio we must find the geometric mean of the given ratios using the formula, $G =$

$$\text{Anti-log} \left(\frac{\Sigma \log x}{n} \right)$$

G denotes geometric mean

x denotes a ratio

$\Sigma \log x$ denotes the sum of log. Values of all the given ratios and n denotes the number of ratios given.

$$n = 2$$

The given ratios are 2 and 4

$$\log 2 = .3010 \quad \log 4 = .6021$$

$$\Sigma \log x = .3010 + .6021 = .9031$$

$$\frac{\Sigma \log x}{n} = \frac{.9031}{2} = .4515$$

Space for hints

$$\text{Anti-log} \left(\frac{\sum \log x}{n} \right) = \text{Anti-log} (.4515) = 2.828$$

$$\therefore G = 2.828$$

Average ratio of wheat to gram prices = 2.828

7. HARMONIC MEAN :

7.1 Definition :

Harmonic Mean (H.M.) is defined as the reciprocal of the arithmetic mean of the reciprocals of the given items.

7.2 Calculation of Harmonic Mean from ungrouped data :

1. Find out the reciprocal of each value in the given set. For example, if 2 is a value, its reciprocal is $1/2$; if 5 is a value its reciprocal is $1/5$.
2. Find out the arithmetic mean of the reciprocals of the given values.
3. Reciprocal of this arithmetic mean is found out next. This gives the value of harmonic mean of the given set of values.

For example, suppose 5, 20, 25, 2, 4 are the given values. To find out the harmonic mean of these values, their reciprocals are found out first.

The reciprocals of the given values are as follows :

$$\frac{1}{5}, \frac{1}{20}, \frac{1}{25}, \frac{1}{2}, \frac{1}{4}$$

$$\text{Arithmetic mean of the reciprocals of given values} = \frac{\frac{1}{5} + \frac{1}{20} + \frac{1}{25} + \frac{1}{2} + \frac{1}{4}}{5}$$

Harmonic mean is the reciprocal of this arithmetic mean

$$\text{In } \frac{\frac{1}{5} + \frac{1}{20} + \frac{1}{25} + \frac{1}{2} + \frac{1}{4}}{5}, \left(\frac{1}{5} + \frac{1}{20} + \frac{1}{25} + \frac{1}{2} + \frac{1}{4} \right) \text{ is called the numerator and}$$

5 is called the denominator. Wherever the reciprocal is needed, the

Check your Progress

13. Define Harmonic Mean.

numerator is written as the denominator and the denominator is written as the numerator and it gives the reciprocal.

Space for hints

$$\frac{5}{\frac{1}{5} + \frac{1}{20} + \frac{1}{25} + \frac{1}{2} + \frac{1}{4}} \text{ is the reciprocal of } \frac{\frac{1}{5} + \frac{1}{20} + \frac{1}{25} + \frac{1}{2} + \frac{1}{4}}{5}$$

$$\therefore \text{ Harmonic mean} = \frac{5}{\frac{1}{5} + \frac{1}{20} + \frac{1}{25} + \frac{1}{2} + \frac{1}{4}} = \frac{5}{.2 + .05 + .04 + .5 + .25} = \frac{5}{1.04} = 4.8$$

In general, the formula to calculate the harmonic mean of a given set of values is as follows,

$$H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_n}}$$

Where H denotes harmonic mean

x_1 denotes the first value in the given set.

x_2 denotes the second value in the given set.

.....

x_n denotes the final value in the given set.

n denotes the total number of values given.

Usually, the sum $\left(\frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_n} \right)$ is denoted by $\sum \frac{1}{x}$. That is, sum of the reciprocals of the given values is denoted by $\sum \frac{1}{x}$.

$$\therefore H = \frac{n}{\sum \frac{1}{x}}$$

Example 1 :

Calculate the harmonic mean of the following set of numbers

80, 90, 120, 60, 96, 85

n = total number of values given = 6

Space for hints

Reciprocals of the given numbers are,

$$\frac{1}{80}, \frac{1}{90}, \frac{1}{120}, \frac{1}{60}, \frac{1}{96}, \frac{1}{85}$$

$$\sum \frac{1}{x} = \text{Sum of the reciprocals of the given numbers}$$

$$= \frac{1}{80} + \frac{1}{90} + \frac{1}{120} + \frac{1}{60} + \frac{1}{96} + \frac{1}{85}$$

$$H = \frac{n}{\sum \frac{1}{x}}$$

$$= \frac{6}{\frac{1}{80} + \frac{1}{90} + \frac{1}{120} + \frac{1}{60} + \frac{1}{96} + \frac{1}{85}}$$

$$= \frac{6}{.0125 + .0111 + .0083 + .0167 + .0104 + .0118}$$

$$= \frac{6}{.0708} = 84.86$$

Answer :

$$\text{Harmonic Mean} = 84.86$$

7.3 Calculation of harmonic mean from discrete frequency distribution :

To calculate the harmonic mean of a discrete frequency distribution.

1. Frequency of each item is divided by the value of the same item. If x is the value of an item and f is the corresponding frequency of the item, f is divided by x and we get $\frac{f}{x}$.

All the values of $\frac{f}{x}$ are found out and summed up. The sum is denoted by $\sum \frac{f}{x}$.

2. All the frequencies are summed up. The sum is denoted by $\sum f$.

Check your Progress

14. Give the formula to calculate H.M. from ungrouped data.

3. Now, the formula to calculate the harmonic mean is given as follows :

Space for hints

$$H = \frac{\sum f}{\sum \frac{f}{x}}$$

Where H denotes harmonic mean

$\sum f$ denotes the sum of frequencies

$\sum \frac{f}{x}$ denotes the sum of all values of $\frac{f}{x}$

Consider the following discrete frequency distribution.

Value of item	Frequency
10	20
20	30
25	50
40	15
50	5

Each frequency is divided by the corresponding value of item and the values of $\frac{f}{x}$ are obtained.

These values are given under $\frac{f}{x}$ in the table below. Their sum gives the value of $\sum \frac{f}{x}$.

Value of item x	Frequency f	$\frac{f}{x}$
10	20	20/10 = 2
20	30	30/20 = 1.5
25	50	50/25 = 2
40	15	15/40 = .375
50	5	5/50 = .1
Total	120	5.975

Space for hints

$$\Sigma f = 120$$

$$\Sigma \frac{f}{x} = 5.975$$

$$H = \frac{\Sigma f}{\Sigma \frac{f}{x}} = \frac{120}{5.975}$$

$$= \frac{24}{1.195} = 20.1 \text{ (approx.)}$$

$$\text{Harmonic mean} = 20.1$$

Example 2 :

Calculate the harmonic mean from the following distribution

Value of item	frequency
.5	4
1.0	15
1.5	20
5.0	16
8.0	9
10.0	5

To calculate the harmonic mean each frequency is divided by the corresponding value of item and given in the column f/x below.

Value of item	Frequency	$\frac{f}{x}$
x	f	
.5	4	$4/.5 = 8$
1.0	15	$15/1 = 15$
1.5	20	$20/1.5 = 13.3$
5.0	16	$16/5 = 3.2$
8.0	9	$9/8 = 1.125$
10.0	5	$5/10 = .5$
Total	69	41.125

$$\Sigma f = 69$$

$$\Sigma \frac{f}{x} = 41.125$$

$$H = \frac{\Sigma f}{\Sigma \frac{f}{x}} = \frac{69}{41.125} = 1.678$$

Example 3 :

Answer : Harmonic mean = 1.678

7.4 Calculation of harmonic mean from continuous frequency distribution :

To Calculate the harmonic mean of continuous frequency distribution :

1. Midvalues of all the classes are found out first.
2. Frequency of each class is divided by the midvalues of the same class. If f is the frequency of a class and x is the midvalue of the same class, then f is divided by x . All the value of $\frac{f}{x}$ are found out and summed up.

The sum is denoted by $\Sigma \frac{f}{x}$.

3. All the frequencies are summed up. The sum is denoted by Σf .
4. The formula to calculate the harmonic mean is given as follows.

$$H = \frac{\Sigma f}{\Sigma \frac{f}{x}}$$

Where H denotes harmonic mean

Σf denotes the sum of frequencies

$\Sigma \frac{f}{x}$ denotes the sum of values of $\frac{f}{x}$.

Consider the following continuous frequency distribution :

Class	Frequency
0 - 10	6
10 - 20	15
20 - 30	18
30 - 40	13
40 - 50	4

Midvalue of each class is found out as follows :

$$\text{Midvalue of the first class} = \frac{0+10}{2} = 5$$

$$\text{Midvalue of the second class} = \frac{10+20}{2} = \frac{30}{2} = 15$$

$$\text{Midvalue of third class} = \frac{20+30}{2} = \frac{50}{2} = 25$$

$$\text{Midvalues of the fourth class} = \frac{30+40}{2} = \frac{70}{2} = 35$$

$$\text{Midvalues of the fifth class} = \frac{40+50}{2} = \frac{90}{2} = 45$$

These midvalues are given under x in the table below.

Then each frequency is divided by the corresponding midvalues which is given under $\frac{f}{x}$ below.

All the values of $\frac{f}{x}$ are summed up and the value of $\Sigma \frac{f}{x}$ is obtained.

All the frequencies are summed up and the value of Σf is obtained.

Midvalue x	Frequency f	$\frac{f}{x}$
5	5	$5/5 = 1$
15	15	$15/15 = 1$
25	18	$18/25 = .72$
35	13	$13/35 = .37$
45	4	$4/45 = .09$
Total	55	3.18

$$\Sigma f = \text{sum of frequencies} = 55$$

$$\sum \frac{f}{x} = 3.18$$

$$H = \frac{\sum f}{\sum \frac{f}{x}} = \frac{55}{3.18} = 17.3 \text{ (approx.)}$$

Space for hints

Example 3 :

Calculate the harmonic mean from the following.

Class interval	Frequency
0 - 4	6
4 - 8	10
8 - 12	12
12 - 16	16
16 - 20	6
20 - 24	4
24 - 28	2

To calculate the harmonic mean the midvalue of each class is found out first. These midvalues are given under x below.

Each frequency is divided by the corresponding midvalues and is given under $\frac{f}{x}$ below. Sum of all the values of $\frac{f}{x}$ gives the value of $\sum \frac{f}{x}$.

Mid value x	Frequency f	$\frac{f}{x}$
2	6	$6/2 = 3.$
6	10	$10/6 = 1.66$
10	12	$12/10 = 1.2$
14	16	$16/14 = 1.14$
18	6	$6/18 = .33$
22	4	$4/22 = .18$
26	2	$2/26 = .08$
Total	56	7.59

$$\Sigma f = \text{sum of frequencies} = 56$$

$$\Sigma \frac{f}{x} = 7.59$$

$$H = \frac{\Sigma f}{\Sigma \frac{f}{x}} = \frac{56}{7.59} = 7.378$$

Answer : Harmonic mean = 7.378

7.5 Use of Harmonic mean :

Harmonic mean is the only appropriate average that can be used in averaging speeds and time.

Consider the following example.

Example :

An aeroplane flies around a square runway each side of which is 100 Kms. The aeroplane covers at a speed of 100 Kms. per hour the first side, 200 Kms. per hour the second side, 300 Kms. per hour the third side and 400 Kms. per hour the fourth side. What is the average speed around the square?

To find out the average speed we must find out harmonic mean. The harmonic mean of the four speeds given above is calculated using the

formula $H = \frac{n}{\Sigma \frac{1}{x}}$.

H denotes harmonic mean

n denotes number of speeds given

x denotes a speed

$\Sigma \frac{1}{x}$ denotes the sum of reciprocals of given speeds.

$$\Sigma \frac{1}{x} = \frac{1}{100} + \frac{1}{200} + \frac{1}{300} + \frac{1}{400}$$

$$n = 4$$

$$\therefore H = \frac{4}{\frac{1}{100} + \frac{1}{200} + \frac{1}{300} + \frac{1}{400}}$$

$$= \frac{4}{.01 + .005 + .0033 + .0025}$$

$$= \frac{4}{.0208} = \frac{1}{.0052} = 192.3 \text{ kms. per hour.}$$

Average speed of the aeroplane = 192.3 kms. per hour.

Example :

Three persons A, B, C can prepare a model in 20, 40 and 25 minutes respectively. What is their average speed?

Here also we find out the harmonic mean of the times taken by A, B, C and this gives their average speed.

We use the formula $H = \frac{n}{\sum \frac{1}{x}}$

$$n = 3$$

$$\sum \frac{1}{x} = \frac{1}{20} + \frac{1}{40} + \frac{1}{25}$$

$$= .05 + .025 + .04 = .115$$

$$H = \frac{3}{.115}$$

$$= 26.1 \text{ minutes (approx.)}$$

Average speed of A, B, C in preparing the model = 26.1 minutes

7.6 Relationship existing among arithmetic mean, geometric mean and harmonic mean:

When we calculate the arithmetic mean and geometric mean of a given set of data, the value of arithmetic mean is invariably greater than or equal to the value of the geometric mean.

When we calculate the geometric mean and harmonic mean of the given set of data, the value of geometric mean is invariably greater than or equal to the value of harmonic mean.

Symbolically, we express the above relationships as follows :

We use the symbol 'A.M.' to denote arithmetic mean 'G.M.' to denote geometric mean and 'H.M.' to denote harmonic mean.

To express that one value is greater than or equal to another value, we use the symbol \geq .

Thus, $A.M. \geq G.M.$ and $G.M. \geq H.M.$

Therefore $A.M. \geq G.M. \geq H.M.$

8. RELATIVE MERITS AND DEMERITS OF VARIOUS AVERAGES

8.1. Arithmetic Mean

(A) Merits :

1. Arithmetic Mean has a rigidly defined formula and nothing is left to guesswork or estimation. Its value is always definite.
2. Arithmetic mean is defined as the sum of all the values of items divided by total number of items. This definition needs no explanation and can be understood easily even by the layman.
3. Arithmetic mean is based on all the given values. Even if the value of a single item is left out, mean cannot be calculated.
4. Arithmetic mean is easily computed. The computation is made still easier by employing the shortcut method.
5. When the separate means of two or more sets are known it is possible to find out the mean of the combined set by a simple formula which we have given earlier. Thus, arithmetic mean lends itself for further algebraic treatment.
6. The value of arithmetic mean does not change much from sample to sample of the same size from a given population. It is the most stable average.

(B) Superiority over other averages :

Arithmetic mean possesses most of the characteristics of an ideal

Check your Progress

15. State the relationship existing among A.M., G.M. and H.M.

average. Arithmetic mean is superior to the other averages in the following respects :

Space for hints

1. To calculate the median from ungrouped data we have a certain formula and we have a different formula to calculate the median from a frequency distribution. Therefore, for the same data when it is ungrouped, we get certain value as the median and when it is grouped into a frequency distribution, we get a different value as the median. For the same data, we can calculate the median using the ogive. The value will be slightly different from both the values which we have obtained above.

In the case of mode, we have three different formulae [viz., Mode = Midvalue of the modal class; $\text{Mode} = l + \frac{cf_2}{f_1 + f_2}$, $\text{Mode} = l + \frac{cd_1}{d_1 + d_2}$ to

calculate the mode from a frequency distribution. Hence we will get different values as the mode for the same frequency distribution.

But in the case of mean, we have fundamentally the same formula viz.,

$$\text{mean} = \frac{\text{Total value of the items given}}{\text{Total number of items given}}$$
 whether the given data are in the form of ungrouped data or in the form of a frequency distribution.

2. If we combine two samples to find the median, we have to rearrange the whole data, if ungrouped, in ascending order of magnitude and then find out the median, If they happen to be frequency distributions, they should have the same class intervals. Then only both the distributions can be combined and we can find out the median. The same type of difficulty lies in the case of mode also to determine the modal value of the combined sample. But in the case of mean, if the separate means of two samples are given, calculation of combined sample mean is made very easy by the help of the formula given earlier.
3. In some cases weightage is to be given to extreme values. In such cases median and mode fail to serve the purpose. For, the median takes into account the items falling in the median class and the mode takes into account only the items falling in the modal class. Only arithmetic mean gives weightage to extreme values also.
4. Even if individual values of items are not known we can find out the value of mean if the total value of items and the number of items are known. It is not the case either with the median or the mode.

(C) Superiority of other averages over arithmetic mean :

Now let us consider how the other averages score over arithmetic mean.

1. If the frequency curve is available, mode can be easily fixed. With the help of the cumulative frequency graphs median can also be easily fixed. But mean cannot be fixed with the help of any graph. It has always to be calculated.
2. Mean gives weightage to extreme values which sometimes results in undesirable effect on the real meaning of an average. For example, suppose 9 students in a class have an average pocket money of one rupee per student while one student has Rs.100 as pocket money. Now, the arithmetic mean for the class as a whole goes up to Rs.10-90. But this value does not reflect the real average obtained in that class. But median and mode do not suffer from this defect.
3. Mean can ignore the value of any single item only at the risk of losing its accuracy. But median and mode can be calculated even when the values of extreme items are not known (as in the case of a distribution having open intervals at its ends).
4. Sometimes we may have attributes which cannot be measured quantitatively. In such cases mean cannot be found out. But we can find the median. For example, we can arrange a group of girls according to their beauty, and select the girl of median beauty whereas it has no arithmetic mean.
5. On some occasions arithmetic mean may give us absurd results, such as 4.6 children born per family, 3.59 rooms per house etc. In such cases, the mode is the best average giving a number for which data actually exists.
6. Arithmetic mean might lead to fallacious conclusions when the actual values from which it is obtained are not given. Suppose two students A and B got the following marks :

	A	B
First terminal examination	45%	75%
Second terminal examination	60%	60%
Annual examination	75%	45%

The arithmetic mean of the marks obtained by both of them are the same viz., 60 But A's progress is positive while B's progress is negative. If along with the average mark viz., 60 their individual marks in the three examinations are not given, one may not know the fact that A is progressing while B is deteriorating. Hence, a fallacious conclusion that the standard of both of them is the same would be drawn.

However, the merits far outweigh the demerits in the case of the mean. Hence, mean should be used wherever possible.

8.2 Median

(A) Merits :

1. Median has a well defined formula.
2. The calculations involved in finding the median are simple and readily understood particularly in the cases of ungrouped data and discrete frequency distribution.
3. In most cases, it is a proper representative of the data in hand and its meaning is easily understood.

(B) Superiority of median over other averages :

1. Median is better than the arithmetic mean in that it is not affected by the values of items on the extremes. If a few extreme values are added to the existing distribution the median moves a little so that it may become the middle most.
2. Median can be calculated without a knowledge of values of extreme items, provided the total number of items is known.
3. Median is more exact in its determination than the mode because, mode cannot be located exactly in a multimode distribution or in a distribution having no mode.
4. If we have a set of 25 students, we can arrange them easily according to increasing order of height and pick out the 13th student as the medianal student. It is enough if we measure his height alone. The height of the other students are immaterial to us whereas they are required for finding out mean or mode.
5. As we have stated earlier, median is specially useful in studying

Check your Progress

16. Which is the best average? Why?

attributes, like beauty, health, colour of hair etc. which are incapable of being quantitatively measured.

Median is highly useful for a correct indication of average wage, wealth, profit etc. In this respect it is better than both mean and mode.

(C) Superiority of other averages over median :

Now let us consider the relative demerits of median.

1. In the case of mean, by multiplying the mean value by the total number of items, we get the sum total of all the values of the items. It is not the case with the median. By multiplying the median by the total number of items, we cannot get the sum total of all the values of items.
2. In case where weightage is to be given to the extreme items, median is less useful than mean though it is better than the mode.
3. Median cannot be precisely determined when it falls between two values; it can only be estimated. When estimated, it may be a value not found in the given set.
4. Median requires the given data, if ungrouped, to be arrayed before it can be determined; it is an operation which involves considerable work.

8.3 Mode

(A) Merits :

1. Mode is easily understood. It has a general precise usage.
2. Mode can be easily located, if the frequency curve is available.

(B) Superiority over other averages :

1. Whereas the mean and median may be merely numerical conceptions, the mode corresponds to a 'reality' in the sense that it gives the value that 'occurs' most often. This makes it the most appropriate average in certain practical situations.
2. When we have a discrete frequency distribution, mode has a better meaning than mean or median.
3. For the wholesale manufacturer who is interested in the type which is usually in demand, the mode is the most suitable average. For example, if a shoemaker finds size nine shoes sells much more than any other

size, he makes that size more in number than the others. He is not interested in finding out the mean or median size shoes.

Space for hints

4. To determine the value of the mode it is not necessary to know the values of extreme items. Only the value of the middle items need be known.

(C) Superiority of other averages over mode :

1. Mode has not got a clear formula and can be got by different methods. The values thus calculated differ from each other.
2. Some distributions may have more than one mode. Choosing between these modal values is difficult. But we always get a single mean or median.
3. Mode rejects all exceptional instances and is, therefore, not useful in those cases where weights are to be given to extreme variations.
4. Modal value multiplied by number of item does not give the sum total of all the values of items.
5. Mode is quite unstable.

8.4 Geometric Mean

(A) Merits :

1. It has a well defined formula based on all items given.
2. It gives less weight to large and more weight to small items than does the arithmetic mean. Hence, it is more appropriate than mean where such a weighting is necessary.
3. Geometric mean is particularly useful in dealing with ratios. It is a highly suitable average for use in index numbers.
4. Geometric mean is capable of further algebraic treatment.

(B) Demerits :

1. Geometric mean cannot be used when any of the values given is zero, or negative, for when a value is zero, the product of all values will be zero and hence the geometric mean will also be zero; when a value is negative the product of all values will become negative and the value of geometric mean would become imaginary.
2. Geometric mean is an average not easily understood by the layman.
3. Geometric mean is not easily obtained as the mean or median. It requires the use of the tables of logarithms for its calculation.

8.5 Harmonic Mean

(A) Merits :

1. Harmonic mean has a well defined formula.
2. It is based on all the values given
3. It gives the largest weight to the smallest values and hence it is valuable where such weighting is desirable.

Harmonic mean is highly useful in averaging rates or speed or prices etc.

(B) Demerits :

1. As in the case of geometric mean, here also the harmonic mean becomes zero when any of the values given is zero.
2. Harmonic mean is not easily understood by the layman.

9. Answers to the Check Your Progress Questions :

- | | |
|------------------|-------------------|
| 1. Refer 2.1 | 9. Refer 5.4 (B) |
| 2. Refer 2.2 | 10. Refer 5.4 (D) |
| 3. Refer 2.3 | 11. Refer 6.1 |
| 4. Refer 2.4 | 12. Refer 6.3 |
| 5. Refer 2.4 (B) | 13. Refer 7.1 |
| 6. Refer 3.1 | 14. Refer 7.2 |
| 7. Refer 4.1 | 15. Refer 7.6 |
| 8. Refer 5.1 | 16. Refer 8.2 |

10. Model questions for guidance :

10 Marks Questions (One Page Answer)

1. What is meant by the central tendency of a frequency distribution? Describe any two measures of the same and compare the for merits.
2. How far do the median and the mode serve as reliable measures of the central tendency?
- 3 Explain the circumstances under which the Harmonic Mean can be taken as a measure of central tendency?

4. What are the special advantages in adopting Arithmetic mean as a measure of central tendency?
5. The Geometric Mean and Harmonic Mean are often used as measures of central tendency. Explain when these measures lead to reliable results.
6. Explain with examples how in certain circumstances mode is a reliable average.
7. Which of the three averages, the arithmetic mean, the median and the mode would you prefer in the following cases. Give reasons.
 - (i) the marks obtained by students in an examination
 - (ii) the runs scored by a batsman
 - (iii) hats manufactured by a firm
8. Write a short note on Geometric Mean.

Space for hints

20 Marks Questions (Three Page Answer)

1. Write an essay on the different averages used in statistics.
2. Calculate the Geometric and Harmonic averages:
18, 48, 96, 148, 76, 042, 0079, 239, 348, 42
3. Define Geometric Mean and Harmonic Mean. Explain their uses and limitations.
4. Compare the mean, the median and the mode.
5. Compare the relative advantages of mean and median as measures of central tendency.
6. Discuss the several measures of central tendency and their relative merits and uses.

Exercise :

- 1) The following table gives the heights of a certain variety of plants. Find the median.

Height (items)	30-39	40-49	50-59	60-69	70-79	80-89	90-99
Frequency	15	46	75	53	40	18	3

Space for hints

- 2) The following table gives the heights of certain plants in a group.

Height (items)	158	160	162	164	166	168	170	172	174
Frequency	3	10	27	40	26	20	9	8	7

Obtain the median.

- 3) Calculate the median from the following table

Marks less than	80	70	60	50	40	30	20	10
Students	100	90	80	60	32	23	13	5

- 4) Compute the median of the marks obtained by 100 students in an examination

Marks	0-10	10-20	20-30	30-40	40-50
No. of Students	12	21	23	34	40

- 5) Find out the median of the following series :

Number of employees Number of factories

50 - 100	35
101 - 150	43
151 - 200	18
201 - 250	10
251 - 300	4

- 6) Compute the median wage from the following data :

Wages in Rs.	30-35	35-40	40-45	45-50	50-55	55-60	60-65	65-70
No. of workers	12	18	22	27	17	23	19	8

- 7) Calculate the quartiles from the following table

Marks less than	80	70	60	50	40	30	20	10
No. of students	100	90	80	60	32	20	13	5

- 8) Calculate mode and quartiles for following data

Profit/shop	0 - 10	10 - 20	20 - 30	30 - 40	40 - 50	50 - 60
No. of shops	12	18	27	20	17	6

9) Compute the mode of the following data

Space for hints

Class interval	155-157	158-160	161-163	164-166	167-169	170-172	173-175	176-178	179-181
Frequency	4	8	26	53	89	62	48	14	6

10) Determine the mode in the following distribution

Class interval	16-17	17-18	18-19	19-20	20-21	21-22	22-23	23-24
Frequency	3	13	23	31	18	9	2	1

11) Calculate the mode in the following distribution.

Marks	0-5	5-10	10-15	15-20	20-25	25-30	30-35
Students	4	6	10	16	12	8	4

12) Calculate the mode from the following :

x	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99
f	148	368	499	648	386	211	172	89	66

13) Calculate the mode from the following frequency distribution.

Class interval	0-5	5-10	10-15	15-20	20-25	25-30	30-35	35-40
Frequency	10	12	17	20	20	18	11	10

14) Compute the modal wage from the following data :

Wages in Rs.	30-35	35-40	40-45	45-50	50-55	55-60	60-65	65-70
Workers	12	18	22	27	17	23	19	8

MEASURES OF DISPERSION, SKEWNESS AND KURTOSIS**Introduction :**

In Unit-2, we explained various averages and stated that they are single numbers used to represent whole mass of data. Is the given average value a true representative of the given set of data? Can we draw reliable conclusions based on it? These are some of the questions for which we want answers before making use of the average value. To assess the reliability of the average, we have certain statistical measures called 'measures of dispersion'. We have both algebraic and graphic measures of dispersion. The meaning and computational procedure of these measures are explained in this Unit-3. Sometimes two or more sets of data (that is, two or more distributions) may have the same average value and the same value for measure of dispersion too. Based on these, we may come to a conclusion that the given distributions are identical with one another. But if we look at the individual values belonging to the different distributions given, they may be different. Therefore, the given distributions are not actually identical with each other. In such situations, we will be protected from drawing wrong conclusions by certain measures called measures of skewness and measures of kurtosis. The meaning and computation of these measures are also explained in this Unit-3.

Unit Objectives:

After studying this unit, you would be able to

- (i) understand the meaning and need for measures of dispersion
- (ii) know the definition and the procedure to calculate the important measures of dispersion namely, Range, Quartile Deviation, Mean Deviation and Standard Deviation.
- (iii) understand the meaning and need for the relative measures of dispersion and also their computational procedure.
- (iv) understand the meaning, types and various measures of skewness
- (v) know the meaning of the Kurtosis and its measurement also.

Unit Structure :**I. MEASURES OF DISPERSION**

1. Need and Meaning of Measures of Dispersion
2. Dispersion - Definition
3. Purposes of measuring dispersion
4. Important measures of dispersion

5. Range
6. Quartile Deviation (Q.D.)
7. Mean Deviation
8. Standard Deviation

II. SKEWNESS

III. KURTOSIS

Answers to the Check your Progress

Model questions for guidance

I. MEASURES OF DISPERSION

1. Need and Meaning of Measures of Dispersion

An average indicates only the central tendency of the data given and does not reveal the entire properties of the data. Hence knowing only the average figure, we cannot make confident statement regarding the data given. Also we cannot make comparisons between two or more sets of data and draw valid conclusions knowing only their average figures. For, two or more sets of data may have identical average values but they may differ in their respective formations. So, further analysis of the data is necessary so that these differences between various sets of data may be studied and accounted for.

Consider the following three series.

	Series A	Series B	Series C
	40	36	1
	40	37	9
	40	38	20
	40	39	30
	40	40	40
	40	41	50
	40	42	60
	40	43	70
	40	44	80
Total	360	360	360
Mean	40	4040	

In the series A, the mean is 40 and the values of all the items are identical with the mean, 40. The items are not at all scattered.

In the series B also the mean is 40. But the values of all the items are not identical with the mean value. However, the difference between the value

of each item and the mean value is not very significant. The minimum value in the series is 36 and the maximum value in the series is 44. Thus, all the items are only slightly scattered around the mean value.

In the series C also the mean value is 40. But the value of each item differs much from the mean value. The minimum value in the series is 1 and the maximum value is 80. The mean value is 40 times the minimum value in the series and half of the maximum value in the series. Thus all the items are widely scattered around the mean value.

From the above, it follows that though the averages are identical, the three series widely differ from each other in their formation. It is evident from the above, that a study of the extent of the scatter of the items around an average may also be made to throw more light on the composition of a series. The name given to this **scatter** is "**Dispersion**". It is also called "**Variation**" or "**Spread**". According to Bowley "**Dispersion is the measure of variation of the items**".

2. Dispersion - Definition

The term "Dispersion" is used in two senses in Statistics.

(i) **Dispersion in a general sense** : Dispersion indicates that the value of items in a series is not uniform. That is, the given items differ in their magnitude. This difference may be great or small. Accordingly, to the dispersion is said to be considerable or insignificant. This is rather a general sense in which this term is used. Suppose, the values of one set of items are found to occur between 100 and 150 and another set of values are found to occur between 100 and 1,000. Here the first set of items is said to have a smaller dispersion than the second set of items.

(ii) **Dispersion in a precise sense** : In this sense, dispersion indicates an absolute or relative measure of the differences of the values of various items from a measure of central tendency computed from those items. The difference between the value of any item and a measure of central tendency is technically called "Deviation". Average of the deviations of the values of various items from their measure of central tendency is called the "Measure of Dispersion".

Mean, Median, Mode, Geometric Mean and Harmonic Mean are called averages of the first order. Since in the calculation of measures of dispersion, we average values derived by the use of the averages of the first order, these measures are called averages of the second order.

3. Purposes of Measuring Dispersion :

Dispersion of a set of values is measured for two basic purposes (i) to measure the reliability of averages and (ii) to serve as a basis for control of the variability itself.

(i) Dispersion as a measure the reliability of averages

Suppose we want to measure the cost of living in a large city. For this, we have to find out the average prices of commodities which are used in general by most of the people in the city because we are considering a large city, each commodity may be sold in a number of shops. It would be a difficult task to contact each and every shop, get the prices of commodities and find out the average prices. So let us select five representative shops for each commodity and find out the average prices. First let us consider kerosene. Suppose the price of kerosene varies from Rs. 1–50 per litre to Rs. 1–60 per litre in the five shops and the mean of the five prices of kerosene be Rs. 1–56 per litre. This mean price differs from the maximum price viz., Rs. 1–60 by 4 paise and from the minimum price viz., Rs. 1.50 by 6 paise. That is, the mean price differs from individual price by a few paise only. Thus, the mean price closely represents the individual price at each of the five shops. Hence, this mean price calculated from a sample of five shops, can be used as a reliable estimate of the mean of prices prevailing in the whole city.

On the other hand, suppose the price of certain type of children's dress varies from Rs. 10 to Rs. 30 in five stores. Suppose, the mean of the prices in these five stores is Rs. 18. Here the mean price differs from the maximum of the five prices viz., Rs. 30 by Rs. 12 and from the minimum of the five prices viz., Rs. 10 by Rs. 8. Thus, the mean price differs from individual price by large quantities. This implies that the sample mean price does not closely represent the individual prices in the five stores. Hence, this sample mean price cannot be used as a reliable estimate of mean of the prices prevailing in the whole city.

The fact that whether a sample average can be used as a reliable estimate of the population average is revealed only by our knowledge about the extent of variation or dispersion of items. Thus, to summarize the facts, in most cases, both an average and a measure of dispersion must be presented.

When dispersion is small, the average is a typical value in the sense that it closely represents the individual values, and it is reliable in the sense that it is a good estimate of the corresponding average of the population.

Check your Progress

1. What do you understand by Dispersion?
2. Why do we calculate measures of dispersion?

On the other hand, when the dispersion is great, the average is not so typical and is highly unreliable to take it as an estimate of the average of the population unless a large number of items have been selected from the population to calculate the average.

(ii) Dispersion as a useful indicator to control variation :

The second basic purpose of measuring dispersion is to determine the nature and causes of variation in order to control the variation itself in matters of health; variations in body temperature; pulse beat and blood pressure are basic guides, to diagnosis. Prescribed treatment is designed to control their variation. In industrial production, efficient operation requires control of variation in the quality of the product. The extent and causes of variation are sought through inspection and quality control programmes. Thus, measures of dispersion are basic guides to the control of causes of variation.

4. Important measures of dispersion

1. Range
2. Quartile deviation
3. Mean deviation
4. Standard deviation and
5. Lorenz curve

Each of the first four measures may be stated in two ways. One method of statement shows the absolute amount of deviation while the other presents the relative amount of deviation. Now let us consider both the ways of stating each measure of dispersion one by one.

5. RANGE

5.1 Meaning and Definition:

Range is the simplest measure of dispersion. Range is defined as the difference between the largest and the smallest items of the given distribution. That is,

$$\text{Range} = \text{value of the largest item} - \text{value of smallest item}$$

Example 1 :

Suppose, the following numbers represent the weights (in lbs.) of 9 students :

80, 100, 85, 90, 110, 120, 150, 135, 140.

Check your Progress

3. Give the names of important measures of dispersion.

4. What is Range?

$$\text{Largest Weight} = 150 \text{ lbs}$$

$$\text{Smallest weight} = 80 \text{ lbs}$$

$$\text{Range} = \text{value of the largest} - \text{the smallest weight}$$

$$= (150 - 80) \text{ lbs}$$

$$= 70 \text{ lbs.}$$

Space for hints

5.2 Absolute measure of Range - Meaning :

Range as calculated above is an absolute measure of dispersion. An absolute measure is always expressed in terms of the unit in which the given distribution is expressed. For instance, in the above illustration, the distribution is given in terms of lbs, and range is also stated in terms of lbs. Hence, the dispersions of two distributions given in two different units cannot be compared with the help of the absolute measure, range.

For example, the range of weight measurements cannot be compared with the range of height measurements since, weights will be in lbs while heights will be in inches.

5.3 Relative measure of Range - Meaning and Need :

For purposes of comparison a relative measure of range is calculated. A relative measure is not expressed in any unit. It is free from units of measurements. It is only a mere number.

The relative measure is called "Coefficient of Dispersion" or ratio of dispersion. It is given as follows :

$$\% \text{ Coefficient of dispersion} = \frac{(\text{Value of the largest item} - \text{Value of the smallest item})}{(\text{Value of the largest item} + \text{Value of the smallest item})}$$

Value of the largest item is denoted by L and value of smallest item is denoted by S.

Check your Progress

5. What do you understand by absolute measures of dispersion?

$$\text{Coefficient of dispersion} = \frac{L-S}{L+S}$$

For the example we have given above,

$$\text{Coefficient of dispersion} = \frac{150-80}{150+80} = \frac{70}{230} = \frac{7}{23} = 0.304$$

Example 2 :

Calculate range and its coefficient for the following data of salaries of 5 persons.

20, 30, 80, 90, 100.

Value of the smallest item = $S = 20$

Value of the largest item = $L = 100$

Range = $L - S = 100 - 20 = 80$ Rupees

$$\text{Coefficient of Range} = \frac{L-S}{L+S}$$

$$= \frac{100-20}{100+20} = \frac{80}{120} = \frac{2}{3}$$

$$= 0.67$$

Example 3 :

Calculate range and its coefficient from the following data.

Marks	Number of students
10 – 20	3
20 – 30	4
30 – 40	5

Value of largest item = $L = 40$

Check your
Progress

6. Define Quartile
Deviation.

Value of smallest item = $S = 10$

Range = $L - S = 40 - 10 = 30$ marks

$$\text{Coefficient of Range} = \frac{L - S}{L + S}$$

$$= \frac{40 - 10}{40 + 10} = \frac{30}{50}$$

$$= \frac{3}{5} = 0.6$$

6. QUARTILE DEVIATION (Q.D.)

6.1 Definition :

A measure of dispersion in terms of the lower and upper quartiles is called "Quartile Deviation" or "Semi-inter-quartile range". It is given as follows.

$$\text{Q.D.} = \frac{Q_3 - Q_1}{2}$$

Where Q.D. denotes quartile deviation.

Q_3 denotes the upper or third quartile.

Q_1 denotes the lower or first quartile.

This is an absolute measure of dispersion. Hence, it is not possible to compare the relative dispersion of two distributions expressed in different units with the help of their quartile deviation.

6.2 Coefficient of Quartile Deviation - Definition

We give below a relative measure of dispersion called quartile coefficient of dispersion or coefficient of quartile deviation.

$$\begin{aligned} \text{Quartile coefficient of dispersion} &= \frac{\frac{Q_3 - Q_1}{2}}{\frac{Q_3 + Q_1}{2}} \\ &= \frac{Q_3 - Q_1}{Q_3 + Q_1} \end{aligned}$$

Space for hints

Check your Progress

7. Give the formula for coefficient of Q.D.

6.3 Computation of Q.D. and coefficient of Q.D.

To calculate both quartile deviation and quartile coefficient of dispersion first of all we must calculate the first quartile (Q_1) and the third quartile (Q_3) from the given set of data. Then by substituting the values of Q_1 and Q_3 in the respective formula given above we get the values of quartile deviation and quartile coefficient of dispersion.

Consider the following examples.

Examples 1 :

Calculate the quartile deviation and the quartile coefficient of dispersion of the following :

31, 37, 43, 48, 57, 60

n = Total number of items given

= 6

$$Q_1 = \text{value of the item} \left(\frac{n+1}{4} \right)$$

$$= \text{Value of the item} \left(\frac{6+1}{4} \right)$$

$$= \text{Value of the item} \left(\frac{7}{4} \right)$$

$$= \text{value of the item } 1.75$$

$$= \text{value of the item } 1 + \frac{3}{4} \text{ (value of the item 2}$$

$$- \text{value of the item 1})$$

$$= 31 + \frac{3}{4} (37 - 31)$$

$$= 31 + (3/4 \times 6)$$

$$= 31 + \frac{9}{2}$$

$$= 31 + 4.5$$

$$= 35.5$$

$$Q_3 = \text{value of the item } \frac{3(n+1)}{4}$$

$$= \text{Value of the item } \frac{3 \times 7}{4}$$

$$= \text{Value of the item } \frac{21}{4}$$

$$= \text{value of the item } 5.25$$

$$= \text{value of the item } 5 + \frac{1}{4} (\text{value of the item 6} - \text{value of the item 5})$$

$$= 57 + \frac{1}{4} (60 - 57)$$

$$= 57 + (\frac{1}{4} \times 3)$$

$$= 57 + \frac{3}{4}$$

$$= 57.75$$

$$\text{Quartile deviation} = \frac{Q_3 - Q_1}{2}$$

$$= \frac{57.75 - 35.5}{2}$$

$$= \frac{22.25}{2}$$

$$= 11.125$$

$$\text{Quartile coefficient of dispersion} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

$$= \frac{57.75 - 35.5}{57.75 + 35.5}$$

$$= \frac{22.25}{93.25}$$

$$= 0.2386.$$

Answer :

$$\text{Quartile deviation} = 11.125$$

$$\text{Quartile coefficient of dispersion} = 0.2386$$

Example 2 :

Calculate the quartile deviation and coefficient of quartile deviation of the following

160, 158, 170, 142, 175

The given items are arranged in ascending order of magnitude as follows :

142, 158, 160, 170, 175

$$n = \text{Total number of items given}$$

$$= 5$$

$$Q_1 = \text{value of the item} \left(\frac{n+1}{4} \right)$$

$$= \text{Value of the item} \left(\frac{5+1}{4} \right)$$

$$= \text{Value of the item} \left(\frac{6}{4} \right)$$

$$= \text{value of the item } 1.5$$

$$= \text{value of the item } 1 + 1/2(\text{value of the item 2} - \text{value of the item 1})$$

$$= 142 + 1/2 (158 - 142)$$

$$= 142 + (1/2 \times 16)$$

$$= 142 + 8$$

$$= 150$$

$$Q_3 = \text{value of the item} \frac{3(n+1)}{4}$$

$$= \text{Value of the item} \frac{3(5+1)}{4}$$

$$= \text{Value of the item } \frac{3 \times 6}{4}$$

$$= \text{Value of the item } \frac{3 \times 6}{4}$$

$$= \text{value of the item } 4.5$$

$$= \text{value of the item } 4 + \frac{1}{2}(\text{value of the item } 5 - \text{value of the item } 4)$$

$$= 170 + \frac{1}{2}(175 - 170)$$

$$= 170 + (\frac{1}{2} \times 5)$$

$$= 170 + 2.5$$

$$= 172.5$$

$$\text{Quartile deviation} = \frac{Q_3 - Q_1}{2}$$

$$= \frac{172.5 - 150}{2}$$

$$= \frac{22.5}{2}$$

$$= 11.25$$

$$\text{Coefficient of quartile deviation} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

$$= \frac{172.5 - 150}{172.5 + 150}$$

$$= \frac{22.5}{322.5}$$

$$= 0.06976$$

Answer :

$$\text{Quartile deviation} = 11.25$$

$$\text{Coefficient of quartile deviation} = 0.06976$$

Example 3 :

The following table gives the heights of certain number of students. Calculate

- (i) First and third quartiles
- (ii) Quartile deviation
- (iii) Quartile coefficient of dispersion

Height in inches	No. of students
58	15
59	20
60	32
61	35
62	33
63	22
64	20

Frequencies are cumulated and the table is given as follows :

Height (inches) x	Frequency (No. of students) f	Cumulative frequency
58	15	15
59	20	35
60	32	67
61	35	102
62	33	135
63	22	157
64	20	177
Total	177	

First we calculate the lower quartile as follows :

$$\begin{aligned}
 N &= \text{Total frequency} \\
 &= 177
 \end{aligned}$$

Example 4 : $\frac{N+1}{4} = \frac{177+1}{4} = \frac{178}{4} = 44.5$

Calculate the quartile deviation and quartile coefficient of dispersion of the following frequency distribution :

$$Q_1 = \text{value of } \left(\frac{N+1}{4} \right)^{\text{th}} \text{ item}$$

$$= \text{value of 44.5th item.}$$

From cumulative frequency column of the table given above we came to know that all the items after the item 35 and upto the item 67 are having their values equal to 60.

The 44.5th item is in between the item 35 and 67

∴ The value of 44.5th item is equal to 60.

(i.e.) The value of Q_1 is equal to 60.

$$Q_1 = 60$$

First quartile = 60 inches.

Now we calculate the upper quartile as follows :

$$\frac{3(N+1)}{4} = \frac{3(N+1)}{4} = \frac{3 \times 178}{4} = \frac{534}{4} = 133.5$$

$$Q_3 = \text{value of } \frac{3(N+1)}{4} \text{ th item.}$$

$$= \text{value of 133.5th item.}$$

From the cumulative frequency column of the table given above we come to know that all the items beyond the item 102 and upto the item 135 are having their values equal to 62.

133.5th item is in between the items 102 and 135

∴ The value of 133.5th item is equal to 62.

$$(i.e.) Q_3 = 62.$$

Third quartile = 62 inches.

Having found out the values of Q_1 and Q_3 we calculate the quartile deviation as follows :

$$\text{Quartile deviation} = \frac{Q_3 - Q_1}{2}$$

$$= \frac{62 - 60}{2} \text{ inches.}$$

$$= \frac{2}{2} \text{ inches.}$$

$$= 1 \text{ inch.}$$

Now, we calculate the quartile coefficient of dispersion as follows :

$$\text{Quartile coefficient of dispersion} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

$$= \frac{62 - 60}{62 + 60}$$

$$= \frac{2}{122}$$

$$= \frac{1}{61}$$

$$= 0.01639^*$$

Answer :

First quartile = 60 inches

Third quartile = 62 inches

Quartile deviation = 1 inch

Quartile coefficient of dispersion = 0.01639

* Note that while quartile deviation is expressed in the units in which the given data are expressed, quartile coefficient of dispersion is a mere number having no units.

Example 4 :

Space for hints

Calculate the quartile deviation and quartile coefficient of dispersion of the following frequency distribution of weights of certain number of persons.

Weight (in lbs)	Frequency
80 – 90	1
90 – 100	11
100 – 110	25
110 – 120	37
120 – 130	15
130 – 140	9
140 – 150	5

The frequencies are cumulated and the frequency tables are given as follows :

Weights (in lbs)	Frequency	Cumulative frequency
80 – 90	1	1
90 – 100	11	12
100 – 110	25	37
110 – 120	37	74
120 – 130	15	89
130 – 140	9	98
140 – 150	5	103
Total	103	

To find out the quartile deviation, we calculate the two quartiles, Q_1 and Q_3 first.

Q_1 is calculated as follows :

$$N = \text{Total frequency}$$

$$= 103$$

Space for hints

$$\frac{N}{4} = \frac{103}{4}$$

$$= 25.75$$

$$Q_1 = \text{value of the item } \left[\frac{N}{4} \right]$$

$$= \text{value of the item } 25.75$$

We come to know from the cumulative frequency column of the table given above that all the items after the item 12 and upto the item 37 have their values in the interval "100 - 110".

The item 25.75 is between the items 12 and 37.

∴ The value of the item 25.75 is in the interval "100 - 110" (i.e.). The value of Q_1 lies in the interval "100 - 110".

∴ "100 - 110" is the Q_1 class

l = True lower limit of the Q_1 class

$$= 100$$

m = cumulative frequency of the class immediately preceding the Q_1 class

= cumulative frequency of the class "90 - 100"

$$= 12$$

f = frequency of the Q_1 class

$$= 25$$

c = magnitude or length of the Q_1 class

$$= 110 - 100$$

$$= 10$$

$$Q_1 = l + \frac{\frac{N}{4} - m}{f} \times c$$

$$= \left(100 + \frac{25.75 - 12}{25} \times 10 \right) \text{ lbs.}$$

$$= 100 + \frac{13.75}{25} \times 10 \text{ lbs.}$$

$$= (100 + 5.50) \text{ lbs} = 105.5 \text{ lbs.}$$

$$\therefore Q_1 = 105.5 \text{ lbs.}$$

Q_3 is calculated as follows :

$$\frac{3N}{4} = \frac{3 \times 103}{4}$$

$$= \frac{309}{4}$$

$$= 77.25$$

$$Q_3 = \text{value of the item } \left(\frac{3N}{4} \right)$$

$$= \text{value of the item } 77.25$$

We come to know from the cumulative frequency column of the above table that all the items after the item 74 and upto the item 89 are having their values in the interval "120 – 130".

The item 77.25 is between the items 64 and 89.

\therefore The value of the item 77.25 is in the interval "120 – 130"

(i.e.,) The value of Q_3 lies in the interval "120 – 130".

\therefore "120 – 130" is the Q_3 class

l = lower limit of the Q_3 class

$$= 120$$

m = cumulative frequency of the class immediately preceding the Q_3 class

= cumulative frequency of the class "110 - 120"

$$= 74$$

f = frequency of the Q_3 class

$$= 15$$

c = magnitude or length of the Q_3 class

$$= 130 - 120$$

$$= 10$$

$$Q_3 = l + \frac{\frac{N}{4} - m}{f} \times c$$

$$= 120 + \left(\frac{77.25 - 74}{15} \times 10 \right) \text{ lbs.}$$

$$= 120 + \left(\frac{3.25}{3} \times 2 \right) \text{ lbs.}$$

$$= 120 + \left(\frac{6.50}{3} \right) \text{ lbs}$$

$$= 120 + 2.166 \text{ lbs}$$

$$= 122.166 \text{ lbs}$$

$$\therefore Q_3 = 122.166 \text{ lbs.}$$

Now the quartile deviation is calculated as follows :

Space for hints

$$\begin{aligned}\text{Quartile deviation} &= \frac{Q_3 - Q_1}{2} \\ &= \frac{122.166 - 105.5}{2} \text{ lbs} \\ &= \frac{16.666}{2} \text{ lbs} \\ &= 8.333 \text{ lbs}\end{aligned}$$

Quartile coefficient of dispersion is calculated as follows :

$$\begin{aligned}\text{Quartile coefficient of dispersion} &= \frac{Q_3 - Q_1}{Q_3 + Q_1} \\ &= \frac{122.166 - 105.5}{122.166 + 105.5} \\ &= \frac{16.666}{227.666} \\ &= 0.0732\end{aligned}$$

Answers :

$$\text{Quartile deviation} = 8.33 \text{ lbs}$$

$$\text{Quartile coefficient of dispersion} = 0.0732$$

6.4 Superiority of Q.D. over Range

The quartiles are more stable compared to the extreme values of a given distribution. Hence, the measure of dispersion based on the quartiles viz., quartile deviation is better than the range.

Check your progress
8. Define Mean Deviation.
9. What do you understand by absolute deviation?

7. MEAN DEVIATION (M.D.)

7.1 Definition

For the data given, first we calculate an average (i.e., mean or median or mode), then we find out the difference between the value of each item in the given data and the average calculated ignoring the signs (in other words, we find out absolute deviation of each item from the average calculated), and finally we calculate the arithmetic mean of these absolute deviations. The value that we get is called mean deviation. Thus, mean deviation can be defined as the arithmetic mean of the absolute deviations of various items from a particular average (either mean or median or mode.)*

7.2 Explanation

Let us consider an example. Suppose marks secured by four students in an examination are 43, 47, 48 and 54. The arithmetic mean for the given data is :

$$\left[\frac{43+47+48+54}{4} = \frac{192}{4} = 48 \right]$$

The difference between the first item in the data given and the arithmetic mean is $= 43 - 48 = -5$, The second item and the arithmetic mean is $= 47 - 48 = -1$, the third item and the arithmetic mean is $= 48 - 48 = 0$, the fourth item and the arithmetic mean is $54 - 48 = +6$. Thus, the deviations are $-5, -1, 0$ and $+6$. If we add these, the value that we get is zero.

(i.e.) The Sum of all the deviations $= -5 - 1 + 6 = -6 + 6 = 0$

To avoid the possibility of getting zero sum of the deviations while calculating mean deviation, we do not consider the sign in front of the numbers that denote the differences between the value of each item in the data given and the average calculated. In other words, we ignore the signs in front of the numbers. We do not bother whether they have positive sign (plus sign) or negative sign (minus sign). By ignoring the signs, we simply take note of the numbers and these numbers are called absolute deviations. In our example, the differences are $-5, -1, 0$ and $+6$. Since we have to ignore the plus or minus signs in front of these differences to calculate mean deviation, we take these differences to be $5, 1, 0$ and 6 for purposes of calculation of mean deviation.^{\$\$}

*Calculation of mean deviation will be very easy if one is well-versed in the calculation of arithmetic mean and median, So the students are advised to go through Unit 2 dealing with arithmetic mean and median once again before they take up this Unit for study.

^{\$\$}If we take note of the plus or minus sign in front of the numbers that denote the differences between the value of each item in the given data the difference is called ordinary deviation. Instead, if we ignore the plus or minus sign in front of the numbers, the difference is called absolute deviation. From what we have explained above it must be clear that what matters in the calculation of mean deviation is absolute deviation and not ordinary deviation. There is an important reason for this. In mean deviation, we are not concerned with knowing whether a large number of items in the data given have greater values than the average calculated or not. Instead, in mean deviation, we are concerned only with measuring the extent of variation or deviation of the items from the average calculated. To serve this purpose absolute deviation will do.

Check your Progress

8. Define Mean Deviation.

9. What do you understand by absolute deviation?

Mean deviation is also called average deviation or Mean variation.

Space for hints

The particular average from which the absolute deviations are taken should be stated specifically. When the absolute deviations are taken from the mean of the given distribution, we call the measure of dispersion as "mean deviation from mean".

If the absolute deviations are taken from the median, we call the measure of dispersion as "mean deviation from median".

If the absolute deviations are taken from the mode we call the measure of dispersion as "mean deviation from mode".

It is not a common practice to calculate the absolute deviations from the mode. In actual practice, mean deviation is calculated either from mean or from median.

7.3 Calculation of mean deviation from ungrouped data

A) Calculation of mean deviation from mean :

We calculate the mean deviation from mean of ungrouped data as follows :

1. First we find out the mean of the given set of data using the formula,

$$\bar{x} = \frac{\sum x}{n}$$

2. The absolute deviation of each item from the mean is found out. We denote the deviation by the symbol $|d|$. This Symbol $|d|$ is read as "modulus d".
3. All the absolute deviations are summed up. The sum of the absolute deviations is denoted by $\Sigma|d|$.
4. Mean deviation from mean is calculated using the following formula viz.,

$$\text{M.D. from mean} = \frac{\Sigma|d|}{n}$$

Where M.D denotes mean deviation.

$|d|$ denotes absolute deviation of an item from mean.

$\Sigma|d|$ denotes the sum of absolute deviations.

n denotes the total number of items given.

Example 1 :

Calculate the mean deviation from mean of the following set of items.
30, 90, 20, 10, 80, 70.

Mean of the above items is calculated first as follows :

$$\Sigma x = \text{sum of the values of given items.}$$

Check your
Progress
10. Give the
formula
to
calculate M.D. for
ungrouped data.

Space for hints

$$= 30 + 90 + 20 + 10 + 80 + 70$$

$$= 300$$

$$n = \text{Total number of items given}$$

$$= 6$$

$$\bar{x} = \frac{\sum x}{n}$$

$$= \frac{300}{6}$$

$$= 50$$

$$\therefore \text{Mean} = 50$$

Now we calculate deviations of the given items from the mean value viz., 50 as follows.

Deviation of the given items from the mean value viz., 50 are $(30 - 50)$, $(90 - 50)$, $(20 - 50)$, $(10 - 50)$, $(80 - 50)$ and $(70 - 50)$.

(i.e.,) the deviations are -20 , 40 , -30 , -40 , 30 and 20 .

By ignoring the sign in front of the above deviations we get the absolute deviations as follows :

20 , 40 , 30 , 40 , 30 and 20 .

Therefore, the values of $|d|$ are 20 , 40 , 30 , 40 , 30 and 20 .

$$\Sigma |d| = \text{sum of the values of } |d|$$

$$= 20 + 40 + 30 + 40 + 30 + 20.$$

$$= 180.$$

$$\text{Now, M.D. from mean} = \frac{\Sigma |d|}{n}$$

$$= \frac{180}{6}$$

$$= 30.$$

$$\therefore \text{Mean deviation from mean} = 30.$$

Check your
Progress

10. Give the formula to calculate M.D. for ungrouped data.

Example 2 :

Space for hints

The following numbers represent the weights (in lbs) of certain number of students. Calculate the mean deviation from mean.

132, 104, 166, 143, 134, 129, 119, 108, 151, 111

The necessary calculation for the computation of M.D from mean may be done as given in the following table :

x (lbs)	$x - \bar{x} = d$	d
132	2.3	2.3
104	-25.7	25.7
166	36.3	36.3
143	13.3	13.3
134	4.3	4.3
129	-0.7	0.7
119	-10.7	10.7
108	-21.7	21.7
151	21.3	21.3
111	-18.7	18.7
1297		155.0

From the table, $\Sigma x = 1297$

$$n = 10$$

$$\bar{x} = \frac{\Sigma x}{n}$$

$$= \frac{1297}{10} = 129.7 \text{ lbs}$$

$$\Sigma |d| = 155$$

$$\text{Now, M.D. from mean} = \frac{\Sigma |d|}{n} = \frac{155}{10}$$

$$= 15.5$$

Mean deviation from mean = 15.5 lbs.

B) Calculation of Mean Deviation from Median :

We calculate the mean deviation from median for ungrouped data as follows :

1. First the given set of items are arrayed and the median is calculated using the formula, Median = value of $\left[\frac{n+1}{2} \right]$ th item.
2. Absolute deviations of the given items from the median are found out. We denote the absolute deviation of an item from the median by $|d|$.
3. All the absolute deviations are summed up. The sum of the absolute deviations is denoted $\Sigma|d|$.
4. Mean deviation from median is calculated by using the following formula viz.,

$$\text{M.D from median} = \frac{\Sigma|d|}{n}$$

Where M.D. denotes mean deviation

$|d|$ denotes absolute deviation of an item from median.

$\Sigma|d|$ denotes sum of the values of $|d|$

n denotes the total number of items given.

Consider the following example.

Example 3 :

Consider the following set of items and calculate the mean deviation from median.

30, 90, 20, 10, 80, 70, 60

First we calculate the median of the given data as follows.

The given set of items is arrayed as follows:

10, 20, 30, 60, 70, 80, 90.

$$n = \text{Total number of items.}$$

$$= 7$$

$$\text{Median} = \text{value of } \left[\frac{n+1}{2} \right] \text{ th item}$$

$$= \text{value of } \frac{7+1}{2} \text{ th item}$$

= value of $\frac{8}{2}$ th item

= value of 4th item.

= 60.

An in Example – 2, here also we can do the calculations in a Table as follows :

x	(x – M) = d	d
10	–50	50
20	–40	40
30	–30	30
60	0	0
70	10	10
80	20	20
90	30	30
		$\Sigma d = 180$

∴ M.D. from median = $\frac{\Sigma|d|}{n}$

= $\frac{180}{7}$

= 25.714 (approx.)

Mean deviation from median = 25.714

7.4 Calculation of Mean Deviation from Discrete Frequency Distribution

(A) Calculation of Mean Deviation from Mean :

We calculate the mean deviation from mean of a discrete frequency distribution as follows :

1. Arithmetic mean of the given distribution is found out using the formula,

$$\bar{x} = \frac{\Sigma xf}{\Sigma f} \text{ or, } \bar{x} = A + \frac{\Sigma fd}{f}$$

2. Absolute deviations of the given items from the mean are found out. The absolute deviation of an item from the mean is denoted by $|d|$.
3. The absolute deviations of each item from mean is multiplied by the frequency of the same item. If f is the frequency of an item whose absolute deviation from mean is $|d|$, then we multiply f and $|d|$. The product we get is $f|d|$.
4. All the values of $f|d|$ are summed up and the sum is denoted by $\Sigma f|d|$.
5. Mean deviation from mean is calculated using the following formula viz.,

$$\text{M.D. from mean} = \frac{\Sigma f|d|}{f}$$

Where M.D. denotes mean deviation.

$|d|$ denotes absolute deviation of an item from mean.

$f|d|$ denotes product of the absolute deviation and corresponding frequency.

$\Sigma f|d|$ denotes the sum of the values of $f|d|$.

Σf denotes the sum of the frequencies.

Consider the following example.

Example 4 :

Calculate the mean deviation from mean of the following discrete frequency distribution.

Marks	No. of Students
10	5
15	10
40	18
50	7
55	4

First we calculate the mean as follows.

Let us choose A to be equal to 40.

Space for hints

Marks x	Frequency f	(x - A) = d	fd
10	6	-30	-180
15	10	-25	-250
40	18	0	0
50	7	10	70
55	4	15	60
Total	45		-300

$$\Sigma f = 45, \quad \Sigma fd = -300$$

$$\begin{aligned}\bar{x} &= A + \frac{\Sigma fd}{f} \\ &= 40 + \frac{-300}{45} \\ &= 40 - \frac{300}{45} \\ &= 40 - 6.66 = 33.34\end{aligned}$$

$$\text{Mean} = 33.34$$

Deviations of the items given from the mean value are (10 - 33.34), (15 - 33.34), (40 - 33.34), (50 - 33.34) and (55 - 33.34)

(i.e.,) the deviations are -23.34, -18.34, 6.66, 16.66 and 21.66

By ignoring the signs before the deviations we get the absolute deviations.

23.34, 18.34, 6.66, 16.66 and 21.66

These are the values of |d|. These values are given under the heading |d| in the table below.

Now each value of |d| is multiplied by the corresponding frequency and given under the heading f|d| in the table below.

Space for hints

x	f	Absolute deviations from the mean, 33.34 d	f d
10	6	23.34	$6 \times 23.34 = 140.04$
15	10	18.34	$10 \times 18.34 = 183.40$
40	18	6.66	$18 \times 6.66 = 119.88$
50	7	16.66	$7 \times 16.66 = 116.62$
55	4	21.66	$4 \times 21.66 = 86.64$
Total	45		646.58

$$\Sigma f = 45 \quad \Sigma f|d| = 646.58$$

$$\begin{aligned} \text{M.D. from mean} &= \frac{\Sigma f|d|}{\Sigma f} \\ &= \frac{646.58}{45} \\ &= \frac{129.316}{9} = 14.368 \end{aligned}$$

∴ Mean deviation from mean = 14.368 marks.

Example 5 :
Calculate the mean deviation from mean of the following distribution

Value of the item	Frequency
4	2
6	4
8	5
10	3
12	2
14	1
16	4

Check your Progress
11. Give the formula to calculate M.D. from discrete frequency distribution.

First we calculate the mean as follows :

Space for hints

Let us assume A to be equal to 10

Value of the item x	Frequency f	(x - A) = d	fd
4	2	-6	-12
6	4	-4	-16
8	5	-2	-10
10	3	0	0
12	2	2	4
14	1	4	4
16	4	6	24
Total	21		-6

$$\Sigma f = 21, \Sigma fd = -6$$

$$\bar{x} = A + \frac{\Sigma fd}{\Sigma f}$$

$$= 10 + \frac{-6}{21} = 10 - \frac{2}{7} = 10 - 0.3(\text{approx.}) = 9.7$$

The absolute deviation i.e., |d| values and each value of |d| multiplied by the corresponding frequency namely, f|d| are given in the following table :

x	f	Absolute deviation d	f d
4	2	5.7	$2 \times 5.7 = 11.4$
6	4	3.7	$4 \times 3.7 = 14.8$
8	5	1.7	$5 \times 1.7 = 8.5$
10	3	0.3	$3 \times 0.3 = 0.9$
12	2	2.3	$2 \times 2.3 = 4.6$
14	1	4.3	$1 \times 4.3 = 4.3$
16	4	6.3	$4 \times 6.3 = 25.2$
Total	21		69.7

$$\Sigma f = 21, \Sigma f |d| = 69.7$$

$$\begin{aligned}\text{M.D. from mean} &= \frac{\sum f|d|}{\sum f} \\ &= \frac{69.7}{21} = 3.32 \text{ (approx.)}\end{aligned}$$

∴ Mean deviation from mean = 3.32

(B) Calculation of mean deviation from median of a discrete frequency distribution :

We calculate the mean deviation from median as follows :

1. We first find out the median of the given distribution using the formula:

$$\text{Median} = \text{value of } \left[\frac{N+1}{2} \right] \text{th item}$$

2. Absolute deviation of each item from median is found out and is denoted by $|d|$.
3. Absolute deviation of each item is multiplied by the frequency of the same item. If f is the frequency of the item whose absolute deviation is $|d|$, then f and $|d|$ are multiplied. The product is $f|d|$.
4. All the values of $f|d|$ are summed up and the sum is denoted by $\sum f|d|$.
5. Mean deviation from median is calculated using the formula

$\sum f$ denotes the sum of frequencies

Example-6:

Consider the same distribution given under Example-5 and calculate the mean deviation from median.

Mark	No. of Students
10	6
15	10
40	18
50	7
55	4

First we calculate the median as follows:

To calculate the median the cumulative frequencies are calculated and the frequency table is given as follows:

Marks	Frequency	Cumulative Frequency
10	6	6
15	10	16
40	18	34
50	7	41
55	4	45
Total	45	

$$N = \text{Total frequency} = 45$$

$$\text{Median} = \text{value of } \left[\frac{N+1}{2} \right] \text{th item}$$

$$= \text{value of } \left[\frac{45+1}{2} \right] \text{th item}$$

$$= \text{value of 23rd item.}$$

We come to know from the cumulative frequency column of the table given above that all the items after the item 16 and upto the item 34 are having their values equal to 40.

The 23rd item is in between the 16th and 34th items.

∴ The value of 23rd item is also equal to 40.

(i.e) Median = 40

Now $|d|$ and $f|d|$ values are as follows:

Marks	Frequency	$ d $	$f d $
10	6	30	$6 \times 30 = 180$
15	10	25	$10 \times 25 = 250$
40	18	0	$18 \times 0 = 0$
50	7	10	$7 \times 10 = 70$
55	4	15	$4 \times 15 = 60$
Total	45		560

$$\Sigma f = 45 \quad \Sigma f|d| = 560$$

$$\text{M.D. from median} = \frac{\Sigma f |d|}{\Sigma f}$$

$$= \frac{560}{45} = 12.44$$

$$\therefore \text{Mean deviation from median} = 12.44$$

M.D. from median always less than M.D. from mean:

In the case of ungrouped data and discrete frequency distribution, the sum of absolute deviations from median is always less than the sum of absolute deviations from mean. This property leads to the fact that mean deviation from median of a given distribution (whether in the form of ungrouped data or discrete frequency distribution) is less than the mean deviation from mean. Thus mean deviation from median gives a more accurate measure of dispersion than the mean deviation from mean.

7.5 Calculation of mean deviation from continuous frequency distribution

(A) Calculation of mean deviation from mean :

We calculate the mean deviation from mean of a continuous frequency distribution as follows

- 1) First we calculate the mean of the continuous frequency distribution

using the formula

Space for hints

$$\bar{x} = \frac{\sum xf}{\sum f} \text{ or } \bar{x} = A + \frac{\sum fd}{\sum f} \times c$$

- 2) Absolute deviation of the midvalue of each class from mean is found out and is denoted by $|d|$.
- 3) Absolute deviation of each class is multiplied by the frequency of the same class. If "f" is the frequency of the item whose absolute deviation is $|d|$ then f and $|d|$ are multiplied. The product is $f|d|$.
- 4) All the values of $f|d|$ are summed up and the sum is denoted by $\sum f|d|$.
- 5) Mean deviation from mean is calculated using the formula.

$$\text{M.D. from mean} = \frac{\sum f|d|}{\sum f}$$

where $\sum f$ denotes the total frequency.

Example 7:

Calculate the mean deviation from mean of the following distribution.

Age (Years)	No. of persons
0-5	7
5-10	10
10-15	16
15-20	32
20-25	24
25-30	18
30-35	10
35-40	5
40-45	1

First we calculate the mean as follows:

Midvalue of each class is found and given under the heading x below.
The midvalue of the 5th class, viz., "20-25" is taken as the value of A .

$$\therefore A = \frac{20+25}{2} = \frac{45}{2} = 22.5$$

Magnitude of each class = 5

$$\therefore c = 5$$

Midvalue x	Frequency f	$d = \frac{x-A}{c}$ $= \frac{x-22.5}{5}$	fd
2.5	7	-4	-28
7.5	10	-3	-30
12.5	16	-2	-32
17.5	32	-1	-32
22.5	24	0	0
27.5	18	1	18
32.5	10	2	20
37.5	5	3	15
42.5	1	4	4
Total	123		-65

$$\bar{x} = A + \frac{\sum fd}{\sum f} \times c$$

$$= 22.5 + \frac{-65}{123} \times 5$$

$$= 22.5 - \frac{325}{123}$$

$$= 22.5 - 2.64$$

$$= 19.86$$

Computation of $|d|$ and $f|d|$ are shown in the following table:

Space for hints

Midvalue x	Frequency f	$ d $	$f d $
2.5	7	17.36	$7 \times 17.36 = 121.52$
7.5	10	12.36	$10 \times 12.36 = 123.60$
12.5	16	7.36	$16 \times 7.36 = 117.76$
17.5	32	2.36	$32 \times 2.36 = 75.52$
22.5	24	2.64	$24 \times 2.64 = 63.36$
27.5	18	7.64	$18 \times 7.64 = 137.52$
32.5	10	12.64	$10 \times 12.64 = 126.40$
37.5	5	17.64	$5 \times 17.64 = 88.20$
42.5	1	22.64	$1 \times 22.64 = 22.64$
Total	123		876.52

$$\Sigma f = 123; \quad \Sigma f|d| = 876.52$$

$$\begin{aligned} \text{M.D. from mean} &= \frac{\Sigma f|d|}{\Sigma f} \\ &= \frac{876.52}{123} \end{aligned}$$

$$= 7.13$$

$$\therefore \text{Mean deviation from mean} = 7.13 \text{ years}$$

(B) Calculation of mean deviation from median

We calculate the mean deviation from median of continuous frequency distribution as follows.

- 1) First we calculate the median of the given distribution using the formula

$$\text{Median} = l + \frac{\frac{N}{2} - m}{f} \times c$$

- 2) The absolute deviation of the midvalue of each class from median is found out and is denoted by $|d|$

- 3) Absolute deviation of the midvalue of each class is multiplied by the frequency of the same class. The product is denoted by $f|d|$.
- 4) All the values of $f|d|$ are summed up and the sum is denoted by $\Sigma f|d|$.
- 5) Mean deviation from median is calculated using the formula,

$$\text{M.D. from median} = \frac{\Sigma f|d|}{\Sigma f}$$

where Σf denotes total frequency.

Example 8:

Consider the table given under Example 10 and calculate the mean deviation from median.

Wages (Rs.)	No. of labourers
20 – 30	3
30 – 40	5
40 – 50	20
50 – 60	10
60 – 70	5

First we calculate the median of the given distribution as follows:

Given frequencies are cumulated and the frequency table is given as follows.

Wages (Rs.)	Frequency	Cumulative Frequency
20 – 30	3	3
30 – 40	5	8
40 – 50	20	28
50 – 60	10	38
60 – 70	5	43
Total	43	

$$N = \text{Total frequency}$$

$$= 43$$

$$\begin{aligned}
 \text{Median} &= \text{value of the item } \left(\frac{N}{2} \right) \\
 &= \text{value of the item } \left(\frac{43}{2} \right) \\
 &= \text{value of the item } 21.5
 \end{aligned}$$

From the cumulative frequency column of the table above we get that all the items after the item 8 and upto the item 28 are having their values in the interval "40-50"

The item 21.5 also is in between items 8 and 28.

∴ The value of the items 21.5 also lies in the interval 40 – 50

(i.e.,) Median lies in the interval "40 – 50"

(i.e.,) "40 – 50" is the median class.

$$\begin{aligned}
 l &= \text{lower limit of the median class.} \\
 &= 40
 \end{aligned}$$

$$\begin{aligned}
 m &= \text{cumulative frequency of the class immediately preceding the median class.} \\
 &= 8
 \end{aligned}$$

$$\begin{aligned}
 f &= \text{frequency of the median class} \\
 &= 20
 \end{aligned}$$

$$\begin{aligned}
 c &= \text{magnitude of the median class} \\
 &= 50 - 40 = 10
 \end{aligned}$$

$$\text{Median} = L + \frac{\frac{N}{2} - m}{f} \times c$$

$$= \text{Rs. } 40 + \frac{21.5 - 8}{20} \times 10$$

$$= \text{Rs. } 40 + \frac{13.5}{2}$$

$$= \text{Rs. } (40 + 6.75)$$

$$= \text{Rs. } 46.75,$$

/d/ and f/d/ values are calculated next as shown in the table below :

Midvalue x	Frequency f	d	f d
25	3	21.75	$3 \times 21.75 = 65.25$
35	5	11.75	$5 \times 11.75 = 58.75$
45	20	1.75	$20 \times 1.75 = 35.00$
55	10	8.25	$10 \times 8.25 = 82.50$
65	5	18.25	$5 \times 18.25 = 91.25$
Total	43		332.75

$$\Sigma f = 43, \Sigma f|d| = 332.75$$

$$\begin{aligned} \text{M.D from median} &= \frac{\Sigma f|d|}{\Sigma f} \\ &= \frac{332.75}{43} = 7.74 \end{aligned}$$

$$\therefore \text{Mean deviation from median} = \text{Rs. } 7.74$$

7.6 Coefficient of Mean Deviation

Mean deviation as calculated above is an absolute measure expressed in the same unit in which the given data are expressed. To enable us to have comparison between two or more distributions expressed in different units we calculate a relative measure of dispersion. This relative measure is called the Mean coefficient of dispersion or the coefficient of M.D. When the deviations are taken from mean.

$$\text{Mean coefficient of dispersion} = \frac{\text{Mean deviation from mean}}{\text{Mean}}$$

When the deviations are taken from median,

$$\text{Mean coefficient of dispersion} = \frac{\text{Mean deviation from median}}{\text{Median}}$$

Example 9 :

Calculate the mean coefficient of dispersion for the data give under Example – 1.

$$\text{Mean} = 50$$

$$\text{Mean deviation from mean} = 30$$

$$\% \text{ Mean coefficient of dispersion} = \frac{\text{Mean deviation from mean}}{\text{Mean}}$$

$$= \frac{30}{50} = 0.6$$

Example 10 :

Calculate the mean coefficient of dispersion for the data give under Example – 3.

We have got the value of median to be 60 and the mean deviation from median to be 25.714

$$\% \text{ Mean coefficient of dispersion} = \frac{\text{Mean deviation from median}}{\text{Median}}$$

$$= \frac{25.714}{60} = 0.4286 \text{ (approx.)}$$

Example 11 :

Find out the mean coefficient of dispersion for the data given under Example –4.

We have found out the value of mean to be 33.34 and the value of mean deviation from mean to be 14.368.

$$\% \text{ Mean coefficient of dispersion} = \frac{14.368}{33.34}$$

$$= 0.431$$

Example 12 :

Calculate the mean coefficient of dispersion for the data given under Example–8.

We have found out the value of median to be 46.75 and mean deviation from median to be Rs. 7.74.

$$\% \text{ Mean coefficient of dispersion} = \frac{7.74}{46.75}$$

$$= 0.1655$$

Note : In the Examination, when you do problems, you need not reproduce all the steps given in the lesson because we have given so many steps in doing problems in the lesson for the sake of your understanding only. It is enough if you give the steps as we have given in the following problems.

Example 13:

Calculate mean deviation* for the following frequency table giving the lengths of 100 telephone calls

Lengths of calls in minutes	0-1	1-2	2-3	3-4	4-5	5-6	6-7	7-8	8-9
Number of calls	12	30	21	16	11	5	2	2	1

Class	Midvalues x	f	fx
0 - 1	0.5	12	6.0
1 - 2	1.5	30	45.0
2 - 3	2.5	21	52.5
3 - 4	3.5	16	56.0
4 - 5	4.5	11	49.5
5 - 6	5.5	5	27.5
6 - 7	6.5	2	13.0
7 - 8	7.5	2	15.0
8 - 9	8.5	1	8.5
Total		100	273.0

$$\Sigma f = 100 \quad \Sigma fx = 273$$

$$\bar{x} = \frac{\Sigma fx}{\Sigma f} = \frac{273}{100} = 2.73$$

Check your Progress

12. How do you calculate M.D. for continuous frequency distribution?

* If you are simply asked to calculate mean deviation it is enough if you calculate mean deviation either from mean or from median alone. You need not calculate both.

x	f	d = [x - \bar{x}]	f d
0.5	12	2.23	26.76
1.5	30	1.23	36.90
2.5	21	0.23	4.83
3.5	16	0.77	12.32
4.5	11	1.77	19.47
5.5	5	2.77	13.85
6.5	2	3.77	7.54
7.5	2	4.77	9.54
8.5	1	5.77	5.77
Total	100		136.98

$$\Sigma f = 100$$

$$\Sigma f|d| = 136.98$$

$$\begin{aligned}
 \text{Mean deviation from mean} &= \frac{\Sigma f|d|}{\Sigma f} \\
 &= \frac{136.98}{100} = 1.3698 \\
 &= 1.37 \text{ (approx.)}
 \end{aligned}$$

Example 14 :

Find the mean deviation for the following frequency distribution with the median as origin.

Length of leaves (Midvalue cm.)	4.0	4.2	4.4	4.6	4.8	5.0	5.2	5.4	5.6	5.8	6.0
Frequency	2	7	10	35	50	90	52	26	12	9	7

Space for hints

Midvalue	Class Interval	Frequency	Cumulative Frequency
4.0	3.9 – 4.1	2	2
4.2	4.1 – 4.3	7	9
4.4	4.3 – 4.5	10	19
4.6	4.5 – 4.7	35	54
4.8	4.7 – 4.9	50	104
5.0	4.9 – 5.1	90	194
5.2	5.1 – 5.3	52	246
5.4	5.3 – 5.5	26	272
5.6	5.5 – 5.7	12	284
5.8	5.7 – 5.9	9	293
6.0	5.9 – 6.1	7	300
Total		300	

$$N = \Sigma f = 300$$

$$\frac{N}{2} = \frac{300}{2} = 150$$

$$\text{Median} = \text{value of the item } \left(\frac{N}{2} \right)$$

$$= \text{value of the item 150}$$

Value of the item 150 falls in the class 4.9 – 5.1

4.9 – 5.1 is the median class

$$l = 4.9$$

$$m = 104$$

$$f = 90$$

$$c = 5.1 - 4.9 = 0.2$$

$$\% \text{ Median} = L + \frac{\frac{N}{2} - m}{f} \times c$$

$$= 4.9 + \frac{150 - 104}{90} \times 0.2$$

$$= 4.9 + \frac{46}{90} \times 0.2$$

$$= 4.9 + \frac{9.2}{90}$$

$$= 4.9 + 0.1$$

$$= 5 \text{ cm.}$$

Midvalue	f	d	f d
4.0	2	1	2.0
4.2	7	0.8	5.6
4.4	10	0.6	6.0
4.6	35	0.4	14.0
4.8	50	0.2	10.0
5.0	90	0	0
5.2	52	0.2	10.4
5.4	26	0.4	10.4
5.6	12	0.6	7.2
5.8	9	0.8	7.2
6.0	7	1	7.0
Total	300		79.8

$$\Sigma f = 300; \quad \Sigma f|d| = 79.8$$

$$\% \text{ Mean deviation from median} = \frac{\Sigma f|d|}{\Sigma f} = \frac{79.8}{300} = 0.266 \text{ cm.}$$

8. STANDARD DEVIATION (S.D.)

8.1 Definition :

Standard Deviation is defined as the square root of the arithmetic mean of the squared deviations of given items from their mean.

In the Calculation of mean deviation we have omitted the signs in front of the deviations so that the sum of the deviations will not be equal to zero but be a positive number. But this operation is mathematically illogical. This drawback is removed in the calculation of standard deviation. Instead of ignoring the signs we can easily make all the deviations positive by squaring them. This process is adopted in the calculation of standard deviation.

8.2 Computation Procedure :

Standard deviation is calculated as follows. First we calculate the arithmetic mean of the given distribution. Next we find out the deviations of various items from the mean calculated. Squares of these deviations are found out. Arithmetic mean of these squared deviations is obtained. Square root of this arithmetic mean gives a value which is called standard deviation.

8.3 Notation used for Standard Deviation

The standard deviation is conventionally represented by the small Greek letter σ (read as sigma). It is also represented by S.D.

The standard deviation is also known as "the root mean square deviation from the mean".

8.4 Variance:

It is defined as the arithmetic mean of the square of the deviations of various items from mean. Hence, variance is nothing but the "square of standard deviation".

$$(i.e.) \text{ variance} = \sigma^2 \quad \text{OR} \quad \sigma = \sqrt{\text{variance}}$$

Variance is also known as "mean square deviation"

Sometimes instead of standard deviation, variance is used to measure dispersion.

8.5 Difference between S.D. and M.D.

Standard deviation is like the mean deviation, a kind of average of all the deviations from mean. But it differs from mean deviation in the following respects:

Check your Progress

13. Define S.D.

14. How do you denote S.D.

15. What is Variance?

16. How is S.D. different from M.D?

1. In the calculation of standard deviation the deviations are always squared and summed up.
2. In the calculation of standard deviation the deviations are always taken from mean; we do not take from median or mode.

8.6 Calculation of standard deviation from ungrouped data:

A) Direct method-1:

We calculate the standard deviation as follows:

1. The arithmetic mean of the data given is found out using the formula,

$$\bar{x} = \frac{\sum x}{n}$$

2. Deviation of each item from mean is found out; if x is the value of an item, its deviation from mean is $(x - \bar{x})$
3. Deviation of each item from mean is squared.

That is, the value of $(x - \bar{x})^2$ is found out.

4. All the squared deviations are summed up. The sum is denoted by $\sum (x - \bar{x})^2$.

5. Total number of items is denoted by n . The sum, $\sum (x - \bar{x})^2$ is divided by n and value of $\frac{\sum (x - \bar{x})^2}{n}$ is found out.

6. Square root of $\frac{\sum (x - \bar{x})^2}{n}$ is found out and this gives the value of standard deviation.

We give the formula to calculate the standard deviation as follows:

$$\sigma = \sqrt{\frac{\sum (x - \bar{x})^2}{n}}$$

Example 1:

Calculate the standard deviation of the following items:

30, 90, 20, 10, 80, 70

Σx = Sum of the values of items given

$$= 30+90+20+10+80+70$$

$$= 300$$

$$n = \text{Total number of items given}$$

$$= 6$$

$$\bar{x} = \frac{\sum x}{n}$$

$$= \frac{300}{6} = 50$$

$$\text{Mean} = 50.$$

Deviations of the given items from mean are $(30-50)$, $(90-50)$, $(20-50)$, $(10-50)$, $(80-50)$, and $(70-50)$,

(i.e.) the deviations are -20 , 40 , -30 , -40 , 30 and 20 .

These are the values of $(x - \bar{x})$. They are given under the heading $(x - \bar{x})$ in the table below.

Squares of the deviations of items from mean are $(-20)^2$, $(40)^2$, $(-30)^2$, $(-40)^2$, $(30)^2$ and $(20)^2$ (i.e.,) 400 , 1600 , 900 , 1600 , 900 and 400

These are the values of $(x - \bar{x})^2$ and they are given under the heading $(x - \bar{x})^2$ in the table below.

$$\begin{aligned} \sum (x - \bar{x})^2 &= \text{Sum of the values of } (x - \bar{x})^2 \\ &= 400+1600+900+1600+900+400 \\ &= 5800. \end{aligned}$$

Now we have given all the values calculated above in a tabular form as follows:

Value of the item (x)	$(x - \bar{x})$	$(x - \bar{x})^2$
30	-20	400
90	40	1600
20	-30	900
10	-40	1600
80	30	900
70	20	400
Total		5800

$$\sigma = \sqrt{\frac{\sum(x - \bar{x})^2}{n}} = \sqrt{\frac{5800}{6}} = \sqrt{\frac{2900}{3}}$$

we find out the square root value using logarithms as follows :

1. We find out the log. value of 2900
2. Next we find out the log-value of 3
3. We subtract the log. value of 3 from the log. value of 2900 and get the balance.

This balance is divided by 2. Anti-log value of the quotient is found out. It is the required square root value.

$$\log 2900 = 3.4624$$

$$\log 3 = 0.4771$$

$$\text{Balance} = 2.9853$$

$$\frac{2.9853}{2} = 1.4926$$

$$\text{Antilog (1.4926)} = 31.10$$

$$\text{That is } \sqrt{\frac{2900}{3}} = 31.1$$

$$\therefore \sigma = \sqrt{\frac{2900}{3}} = 31.1$$

$$\therefore \text{Standard deviation} = 31.1$$

Example 2 :

Find out the standard deviation of the following data

160, 158, 170, 142, 175.

The necessary computations for finding out standard deviation are done in the following table.

Value of item x	$(x - \bar{x})$ $= (x - 161)$	$(x - \bar{x})^2$ $= (x - 161)^2$
160	-1	$(-1)^2 = 1$
158	-3	$(-3)^2 = 9$
170	9	$(9)^2 = 81$
142	-19	$(-19)^2 = 361$
175	14	$(14)^2 = 196$
Total	805	= 648

$$\Sigma x = 805$$

$$n = 5$$

$$\begin{aligned} \bar{x} &= \frac{\Sigma x}{n} \\ &= \frac{805}{5} = 161 \end{aligned}$$

$$\Sigma (x - \bar{x})^2 = 648$$

$$\begin{aligned} \sigma &= \sqrt{\frac{\Sigma (x - \bar{x})^2}{n}} \\ &= \sqrt{\frac{648}{5}} \\ &= \sqrt{129.6} \end{aligned}$$

As we have done in the previous example, we find out the value of 129.6 using log. tables as follows:

1. Find out the log value of 129.6.
2. Divide the above log. value by the index of the root viz., 2.
3. Find out the anti-log. value of the quotient

This gives the value of $\sqrt{129.6}$

Space for hints

$$\log. 129.6 = 2.1127$$

$$\log \frac{129.6}{2} = \frac{2.1127}{2}$$

$$= 1.0563$$

$$\text{Anti-log } (1.0563) = 11.39$$

$$\therefore \sqrt{129.6} = 11.39$$

$$\therefore \sigma = \sqrt{129.6}$$

$$= 11.39$$

$$\therefore \text{Standard deviation} = 11.39$$

B) Direct Method-2 :

We explain this method as follows:

1. Square of the value of each item is found out. If x is the value of an item, value of x^2 is found out.
2. All the values of x^2 are summed up and the sum is denoted by Σx^2 .
3. The values of given items are summed up and the sum is denoted by Σx .
4. We denote the total number of items given by " n ". Now we calculate the standard deviation using the formula.

$$\sigma = \sqrt{\frac{\Sigma x^2}{n} - \left[\frac{\Sigma x}{n} \right]^2}$$

Example 3 :

Calculate the standard deviation of the data given under Example-1 using the formula given above.

30, 90, 20, 10, 80, 70.

Squares of all the values given are found out and given under the

Space for hints

heading x^2 in the table below.

	Value of item x	x^2
	30	$(30)^2 = 900$
	90	$(90)^2 = 8100$
	20	$(20)^2 = 400$
	10	$(10)^2 = 100$
	80	$(80)^2 = 6400$
	70	$(70)^2 = 4900$
Total	300	20800

$$\Sigma x = 300$$

$$\Sigma x^2 = 20800$$

n = Total number of items given

$$= 6$$

$$\sigma = \sqrt{\frac{\Sigma x^2}{n} - \left[\frac{\Sigma x}{n}\right]^2}$$

$$= \sqrt{\frac{20800}{6} - \left[\frac{300}{6}\right]^2}$$

$$= \sqrt{\frac{10400}{3} - (50)^2}$$

$$= \sqrt{\frac{10400}{3} - 2500}$$

$$= \sqrt{\frac{10400 - 7500}{3}}$$

$$= \sqrt{\frac{2900}{3}}$$

As we have found out already in Example 1.

$$\sqrt{\frac{2900}{3}} = 31.1$$

$$\therefore \sigma = 31.1$$

$$\text{Standard deviation} = 31.1$$

(C) Short-cut Method:

We calculate the standard deviation by the short-cut method as follows:

1. We select some value within the range of the given values as the assumed average.

We denote this value by A.

2. Deviation of each item from the assumed average viz. A is found out. If x is the value of an item, the value of (x-A) is found out. We denote the value of (x-A) by the letter d.

3. All the values of d are summed up and the sum is denoted by $\sum d$.

4. Square of each value of d is found out. That is, the value of d^2 are found out.

5. All the values of d^2 are summed up and the sum is denoted by $\sum d^2$.

6. Total number of items given is denoted by n.

Now we calculate the standard deviation using the formula.

$$\sigma = \sqrt{\frac{\sum d^2}{n} - \left[\frac{\sum d}{n} \right]^2}$$

Example 4 :

Calculate the standard deviation of the data given under Example-1 by the short-cut method.

30,90,20,10,80,70.

The given values range between 10 and 90.

∴ We choose some value in between 10 and 90 as the assumed average. Let us choose 60 as the assumed average.

∴ A = 60.

Deviation of each value of x from A is found out and given under the heading $(x-A) = d$ in the table below.

All the values of d are squared and given under the heading d^2 in the table below.

Value of item x	$(x-A) = d$	d^2
30	-30	$(-30)^2 = 900$
90	30	$(30)^2 = 900$
20	-40	$(-40)^2 = 1600$
10	-50	$(-50)^2 = 2500$
80	20	$(20)^2 = 400$
70	10	$(10)^2 = 100$
Total	-60	6400

$$\Sigma d = -60$$

$$\Sigma d^2 = 6,400$$

$$n = \text{Total number of items given.}$$

$$= 6$$

$$\sigma = \sqrt{\frac{\Sigma d^2}{n} - \left[\frac{\Sigma d}{n}\right]^2}$$

$$\sigma = \sqrt{\frac{6400}{6} - \left[\frac{-60}{6}\right]^2}$$

$$= \sqrt{\frac{3200}{3} - (-10)^2}$$

$$= \sqrt{\frac{3200}{3} - 100}$$

$$= \sqrt{\frac{3200 - 300}{3}}$$

$$= \sqrt{\frac{2900}{3}}$$

We have found out the value of $\sqrt{\frac{2900}{3}}$ in Example-1 to be 31.1

$$\sigma = 31.1$$

$$\therefore \text{Standard deviation} = 31.1$$

Example 5 :

Space for hints

Calculate the standard deviation of the data given under Example-2 by the short cut method.

160,158,170,142,175.

The given values range from 142 to 175.

Let us take the value of A to be 150

Deviation of items from 150 is found out and given under the heading d in the table below:

Each value of d is squared and given under the heading d^2 in the table below.

Value of item	d	d^2
160	10	$(10)^2 = 100$
158	8	$(8)^2 = 64$
170	20	$(20)^2 = 400$
142	-8	$(-8)^2 = 64$
175	25	$(25)^2 = 625$
Total	55	1253

$$\Sigma d = 55$$

$$\Sigma d^2 = 1253$$

n = Total number of items given

$$= 5$$

$$\sigma = \sqrt{\frac{\Sigma d^2}{n} - \left(\frac{\Sigma d}{n}\right)^2}$$

$$= \sqrt{\frac{1253}{5} - \left[\frac{55}{5}\right]^2} = \sqrt{250.6 - (11)^2} = \sqrt{250.6 - 121}$$

$$= \sqrt{129.6}$$

In Example-2, we have found out the value of $\sqrt{129.6}$ to be 11.39

$$\sigma = 11.39$$

Standard Deviation = 11.39

8.7 Calculation of standard deviation from discrete frequency distribution

A. Direct method

We calculate the standard deviation by the direct method as follows:

1. We first find out the mean of the given distribution using the formula,

$$\bar{x} = \frac{\sum xf}{\sum f}$$

2. Deviation of each item from the mean value is found out. If x is the value of an item, its deviation from mean is $(x - \bar{x})$

3. Square of each deviation is found out, If $(x - \bar{x})$ is the deviation, the value of $(x - \bar{x})^2$ is found out.

4. Square of the deviation of each item is multiplied by the frequency of the same item. If f is the frequency of the item whose deviation from mean is $(x - \bar{x})$, then f and $(x - \bar{x})^2$ are multiplied. The product which we obtain is $f(x - \bar{x})^2$

5. All the values of $f(x - \bar{x})^2$ are summed up and the sum is denoted by $\sum f(x - \bar{x})^2$.

6. Given frequencies are summed up and the sum is denoted by N . That is, $N = \sum f$

7. Now we calculate the standard deviation using the formula,

$$\sigma = \sqrt{\frac{\sum f(x - \bar{x})^2}{N}}$$

Example 6 :

Calculate the standard deviation for the following discrete frequency distribution.

Marks	No. of Students
20	8
30	12
40	20
50	10
60	6
70	4

3. We first calculate the mean of the given distribution as follows:

Space for hints

	Marks x	No. of students f	fx
	20	8	160
	30	12	360
	40	20	800
	50	10	500
	60	6	360
	70	4	280
Total		60	2460

$$\Sigma f = 60,$$

$$\Sigma xf = 2460$$

$$\bar{x} = \frac{\Sigma xf}{\Sigma f}$$

$$= \frac{2460}{60} = 41$$

Deviation of each item from mean is found out and given under $(x - \bar{x})$ in the table below. Square of each value of $(x - \bar{x})$ is found out and given under the heading $(x - \bar{x})^2$ in table below. Each value of $(x - \bar{x})^2$ is multiplied by the corresponding frequency and given under the heading $f(x - \bar{x})^2$.

$$\bar{x} = 41$$

\bar{x}	f	$(x - \bar{x})$	$(x - \bar{x})^2$	$f(x - \bar{x})^2$
20	8	$20 - 41 = -21$	$(-21)^2 = 441$	$8 \times 441 = 3528$
30	12	$30 - 41 = -11$	$(-11)^2 = 121$	$12 \times 121 = 1452$
40	20	$40 - 41 = -1$	$(-1)^2 = 1$	$20 \times 1 = 20$
50	10	$50 - 41 = 9$	$(9)^2 = 81$	$10 \times 81 = 810$
60	6	$60 - 41 = 19$	$(19)^2 = 361$	$6 \times 361 = 2166$
70	4	$70 - 41 = 29$	$(29)^2 = 841$	$4 \times 841 = 3364$
Total	60			11340

$$N = \text{Total frequency}$$

$$= 60$$

$$\Sigma f(x - \bar{x})^2 = 11340$$

$$\sigma = \sqrt{\frac{\Sigma f(x - \bar{x})^2}{N}}$$

$$= \sqrt{\frac{11340}{60}} = \sqrt{189}$$

We find out the value of $\sqrt{189}$ using logarithms as follows:

$$\log \sqrt{189} = \frac{\log 189}{2}$$

$$= \frac{2.2765}{2}$$

$$= 1.1383$$

$$\text{Anti-log } (1.1383) = 13.75$$

$$\therefore \sqrt{189} = 13.75$$

$$\sigma = 13.75$$

$$\text{Standard deviation} = 13.75$$

B. Short-cut method:

We calculate the Standard deviation by the short-cut method as follows:

1. We take some value (within the range of the values given as the assumed average. We denote the assumed average by the letter. "A".
2. Deviation of each item x from A is calculated and it is denoted by "d".
3. Deviation of each item is multiplied by the frequency of the same item and the products are summed up. The sum is denoted by Σfd .
4. Square of the deviation of each item from A is obtained. If d is the deviation of an item, value of d^2 is found out.
5. Square of the deviation of each item is multiplied by the frequency of the same item. If f is the frequency of an item whose square of the deviation is d^2 , then f and d^2 are multiplied. The product we get is fd^2 .
6. All the values of fd^2 are summed up and the sum is denoted by Σfd^2 .
7. Total value of frequencies is found out and is denoted by N .

8. Now we calculate the standard deviation using the formula

Space for hints

$$\sigma = \sqrt{\frac{\sum fd^2}{N} - \left[\frac{\sum fd}{N} \right]^2}$$

Example 7:

Calculate the standard deviation of the following distribution by the short-cut method.

Marks	No. of Students
10	6
15	10
40	18
50	7
55	4

Let us take A to be 35

Deviation of each item from A viz., 35 is found out and given under the heading d in the table below.

Each value of d is multiplied by the corresponding frequency and given under the heading fd.

Squares of the values of d are found out and given under the heading d² in the table below.

Each value of d² is multiplied by the corresponding frequency and given under the heading fd² in the table below.

A = 35

x	f	d	fd	d ²	fd ²
10	6	-25	6×(-25) = -150	(-25) ² = 625	6×625 = 3750
15	10	-20	10×(-20) = -200	(-20) ² = 400	10×400 = 4000
40	18	5	18×5 = 90	(5) ² = 25	18×25 = 450
50	7	15	7×15 = 105	(15) ² = 225	7×225 = 1575
55	4	20	4×20 = 80	(20) ² = 400	4×400 = 1600
Total	45		-75		11375

Space for hints

N = Total frequency

$$= 45$$

$$\Sigma fd = -75$$

$$\Sigma fd^2 = 11375$$

$$\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left[\frac{\Sigma fd}{N}\right]^2}$$

$$= \sqrt{\frac{11375}{45} - \left[\frac{-75}{45}\right]^2}$$

$$= \sqrt{\frac{2275}{9} - \left[\frac{-5}{3}\right]^2}$$

$$= \sqrt{\frac{2275}{9} - \frac{25}{9}}$$

$$= \sqrt{\frac{2250}{9}}$$

$$= \sqrt{250}$$

$$= \sqrt{250}$$

$$= \sqrt{5 \times 5 \times 10}$$

$$= 5 \times \sqrt{10}$$

$$\frac{\log 10}{2} = \frac{1}{2} = .5$$

$$\text{Antilog}.5 = 3.162$$

$$\therefore \sqrt{10} = 3.162$$

$$\therefore \sigma = 5 \times 3.162$$

$$= 15.81$$

x	f	fd	fd ²
10	6	-60	-360
12	10	-120	-1440
18	22	-396	-7128
20	7	-140	-2800
25	4	-100	-2500
Total	45	-75	11375

Example 8 :

Space for hints

Calculate the standard deviation of the distribution given below by the short-cut method.

Value of the item	Frequency
4	2
6	4
8	5
10	3
12	2
14	1
16	3

Let us take A to be 10

Deviation of each item from A viz., 10 is found out and given under the heading d in the table below.

Each value of d is multiplied by the corresponding frequency and given under the heading fd in the table below.

Squares of the values of d are found out and given under the heading d^2 in the table below.

$$A = 10$$

x	f	d	fd	d^2	fd^2
4	2	-6	$2 \times (-6) = -12$	$(-6)^2 = 36$	$2 \times 36 = 72$
6	4	-4	$4 \times (-4) = -16$	$(-4)^2 = 16$	$4 \times 16 = 64$
8	5	-2	$5 \times (-2) = -10$	$(-2)^2 = 4$	$5 \times 4 = 20$
10	3	0	$3 \times 0 = 0$	$(0)^2 = 0$	$3 \times 0 = 0$
12	2	2	$2 \times 2 = 4$	$(2)^2 = 4$	$2 \times 4 = 8$
14	1	4	$1 \times 4 = 4$	$(4)^2 = 16$	$1 \times 16 = 16$
16	3	6	$3 \times 6 = 18$	$(6)^2 = 36$	$3 \times 36 = 108$
Total	20		-12		288

Space for hints

$N =$ Total frequency

Example 8 :

Calculate the standard deviation given below by the short-cut method.

	$\Sigma fd = -12$
	$\Sigma fd^2 = 288$
	$\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left[\frac{\Sigma fd}{N}\right]^2}$
	$= \sqrt{\frac{288}{20} - \left[\frac{-12}{20}\right]^2}$
	$= \sqrt{\frac{72}{5} - \frac{9}{25}}$

Deviation of each item from 10 is found out and given under the heading d in the table below.

Each value of d is multiplied by the corresponding frequency and given under the heading fd in the table below.

Squares of the values of d are found out and given under the heading fd^2 in the table below.

The value of $\sqrt{351}$ is obtained by using the log table as follows:

$$\log 351 = 2.5453$$

$$\frac{\log 351}{2} = \frac{2.5453}{2} = 1.2726$$

$$\text{Anti-log } (1.2726) = 18.74$$

$$\sqrt{351} = 18.74$$

$$\sigma = \frac{\sqrt{351}}{5}$$

$$= \frac{18.74}{5} = 3.748$$

$$\text{Standard deviation} = 3.748$$

x	f	d	fd	fd^2
4	2	-6	-12	36
6	4	-4	-16	16
8	2	-2	-4	4
10	3	0	0	0
12	2	2	4	4
14	1	4	4	16
16	3	6	18	36
Total	20		-12	112

8.8 Calculation of standard deviation from continuous frequency distribution

Space for hints

(A) Direct Method:

We calculate the standard deviation by the direct method as follows:

1. Midvalue of each class in the given distribution is found out.
2. Mean of the given distribution is calculated using the formula.

$$\bar{x} = \frac{\sum xf}{\sum f} \text{ or } \bar{x} = A + \frac{\sum fd}{\sum f} \times c$$

3. Deviation of the midvalue of each class from mean is found out. If x is the midvalue of a class its deviation from mean is $(x - \bar{x})$.
4. Square of the value of each deviation is found out if $(x - \bar{x})$ is the deviation, value of $(x - \bar{x})^2$ is found out.
5. Square of each deviation is multiplied by the corresponding frequency. These products are summed up and sum is denoted by $\sum f(x - \bar{x})^2$.
6. Total frequency is calculated and is denoted by N .
7. Now we calculate the standard deviation using the formula,

$$\sigma = \sqrt{\frac{\sum f(x - \bar{x})^2}{N}}$$

Example 9:

Calculate the standard deviation of the following distribution

Class	Frequency
0 - 6	4
6 - 12	8
12 - 18	14
18 - 24	15
24 - 30	19

Midvalue of each class is found out and mean is calculated as follows:

Space for hints

Midvalue x	Frequency f	xf
3	4	12
9	8	72
15	14	210
21	15	315
27	19	513
Total	60	1122

$$\sum f = \text{Total frequency}$$

$$= 60$$

$$\sum xf = 1122$$

$$\bar{x} = \frac{\sum xf}{\sum f}$$

$$= \frac{1122}{60} = 18.7$$

Now the deviations of the midvalues of all the classes from mean are (3-18.7), (9-18.7), (15-18.7), (21-18.7), (27-18.7).

(i.e.,) the deviations are -15.7, -9.7, -3.7, 2.3 and 8.3

These are the values of $(x - \bar{x})$ and therefore given under the heading $(x - \bar{x})$ in the table below.

Squares of the deviations are $(-15.7)^2$, $(-9.7)^2$, $(-3.7)^2$, $(2.3)^2$ and $(8.3)^2$.

(i.e.,) the squares of the deviations are 246.49, 94.09, 13.69, 5.29 and 68.89.

These are the values of $(x - \bar{x})^2$ and hence they are given under the heading $(x - \bar{x})^2$ in the table below.

Each value of $(x - \bar{x})^2$ is multiplied by the corresponding frequency as follows:

$$246.49 \times 4 = 985.96$$

$$94.09 \times 8 = 752.72$$

$$13.69 \times 14 = 191.66$$

$$5.29 \times 15 = 79.35$$

$$68.89 \times 19 = 1309.00$$

The above products are the values of $(x - \bar{x})^2$ and hence they are given under the heading $(x - \bar{x})^2$ in the table below.

Midvalue x	f	$(x - \bar{x})$	$(x - \bar{x})^2$	$f(x - \bar{x})^2$
3	4	-15.7	246.49	985.96
9	8	-9.7	94.09	752.72
15	14	-3.7	13.69	191.66
21	15	2.3	5.29	79.35
27	19	8.3	68.89	1309.00
Total		60		3318.69

$$N = \text{Total frequency}$$

$$= 60$$

$$\sum f(x - \bar{x})^2 = \text{sum of all the values of } f(x - \bar{x})^2 = 3318.69$$

$$\therefore \sigma = \sqrt{\frac{\sum f(x - \bar{x})^2}{N}}$$

$$= \sqrt{\frac{3318.69}{60}} = \sqrt{55.3115}$$

We find out the value of $\sqrt{55.3115}$ using log tables as follows:

$$\log 55.3115 = 1.7428$$

$$1/2 \times \log 55.3115 = \frac{1.7428}{2} = .8714$$

$$\text{Anti-log } (.8714) = 7.437$$

$$\sqrt{55.3115} = 7.437$$

$$\sigma = \sqrt{55.311} = 7.437$$

$$\text{Standard deviation} = 7.437.$$

B) Short-cut method:

Standard deviation is calculated by the short-cut method as follows:

1. Midvalue of each class is found out.

2. Midvalue of some class (preferably the midvalue of that class in the centre of the given distribution) is taken as the assumed average. It is denoted by A.
3. Deviation of each midvalue from A is found out. If x is the midvalue of a class its deviation from A is $(x - A)$

4. We divide the deviation of each midvalue from A by the magnitude of the class interval. We denote the magnitude of the class interval by c.

Therefore, if $(x - A)$ is the deviation, it is divided by c and we get the value of $\frac{(x - A)}{c}$. We denote $\frac{(x - A)}{c}$ by d (i.e.) $d = \frac{(x - A)}{c}$.

5. Each value of d is multiplied by the corresponding frequency and the product is denoted by fd.

All the values of fd are summed up and the sum is denoted by $\sum fd$.

6. Square of each value of d is found out. That is, values of d^2 are found out.

7. Each value of d^2 is multiplied by the corresponding frequency and the product is denoted by fd^2 .

All the values of fd^2 are summed up and the sum is denoted by $\sum fd^2$.

8. All the frequencies are summed up and the sum is denoted by N, that is $N = \sum f$

9. Now we calculate the standard deviation using the formula,

$$\sigma = \sqrt{\frac{\sum fd^2}{N} - \left[\frac{\sum fd}{N} \right]^2} \times c$$

Example 10:

Calculate the standard deviation of the following distribution by the short cut method.

Class	Frequency
0 - 6	3
6 - 12	8
12 - 18	14
18 - 24	15
24 - 30	19

Magnitude of the class = 6

Space for hints

$$c = 6$$

Midvalue of each class is found out and given under the heading x in the table below:

$$\text{Midvalue of the third class is } = \frac{12+18}{2} = \frac{30}{2} = 15$$

This value 15 is taken as the assumed average

$$A = 15.$$

Now the deviation of each midvalue from A viz., 15 is found out and divided by the value of c viz., 6.

These are the values of d and given under the heading $d = \left(\frac{x-A}{c} \right)$ in the table below.

Each value of d is multiplied by the corresponding frequency and given under the heading fd in the table below.

Square of each value of d is found out and given under the heading d^2 in the table below.

Each value of d^2 is multiplied by the corresponding frequency and given under the heading fd^2 in the table below:

x	f	$\frac{x-A}{c} = d$	fd	d^2	fd^2
3	4	$\frac{3-15}{6} = -2$	$4 \times (-2) = -8$	$(-2)^2 = 4$	$4 \times 4 = 16$
9	8	$\frac{9-15}{6} = -1$	$8 \times (-1) = -8$	$(-1)^2 = 1$	$8 \times 1 = 8$
15	14	$\frac{15-15}{6} = 0$	$14 \times 0 = 0$	$(0)^2 = 0$	$14 \times 0 = 0$
21	15	$\frac{21-15}{6} = 1$	$15 \times 1 = 15$	$(1)^2 = 1$	$15 \times 1 = 15$
27	19	$\frac{27-15}{6} = 2$	$19 \times 2 = 38$	$(2)^2 = 4$	$19 \times 4 = 76$
Total	60		37		115

Space for hints

$$N = \text{Total frequency} = 60$$

$$\sum fd = 37$$

$$\sum fd^2 = 115$$

$$\sigma = \sqrt{\frac{\sum fd^2}{N} - \left[\frac{\sum fd}{N}\right]^2} \times c$$

$$= \sqrt{\frac{115}{60} - \left[\frac{37}{60}\right]^2} \times 6$$

$$= \sqrt{\frac{115}{60} - \frac{1369}{3600}} \times 6$$

$$= \sqrt{\frac{6900 - 1369}{3600}} \times 6$$

$$= \sqrt{\frac{5531}{3600}} \times 6$$

$$= \sqrt{\frac{5531}{60 \times 60}} \times 6$$

$$= \frac{\sqrt{5531}}{60} \times 6$$

$$= \frac{\sqrt{5531}}{10}$$

Value of $\sqrt{5531}$ is found out using the log-tables as follows:

$$\log. 5531 = 3.7428$$

$$\log \sqrt{5531} = \frac{3.7428}{2} = 1.8714$$

$$\text{Anti-log } 1.8714 = 74.37$$

$$\sqrt{5531} = 74.37$$

$$\therefore \sigma = \frac{\sqrt{5531}}{10} = \frac{74.37}{10} = 7.437$$

$$\text{Standard deviation} = 7.437$$

Example 11:

Space for hints

Calculate the standard deviation from the following data by the short-cut method.

Profits Rs.	No. of firms
0 – 1000	4
1000 – 2000	6
2000 – 3000	10
3000 – 4000	30
4000 – 5000	15
5000 – 6000	10

Midvalue of each class is found out and given under the heading x below.

There are two central classes viz., the third and the fourth classes. Midvalue of anyone of these classes can be taken as the value of A . Let us take the midvalue of the fourth class as the value of A .

$$A = \frac{3000 + 4000}{2} = \frac{7000}{2} = 3500$$

∴ Magnitude of each class = 1000

$$c = 1000$$

Deviation of each midvalue from A viz., 3500 is found out and divided by the value of c viz., 1000.

That is, the values of $\frac{x - A}{c} = d$ are found out. These values are given under the heading $\frac{x - A}{c} = d$ in the table below.

Each value of d is multiplied by the corresponding frequency and given under the heading fd in the table below.

Square of each value of d is found out and given under the heading d^2 in table below.

Each value of d^2 is multiplied by the corresponding frequency given under the heading $f d^2$ in the table below.

$$A = 3500; c = 1000$$

Space for hints

x	f	$\frac{x - A}{c} = d$	fd	d ²	fd ²
500	4	$\frac{500 - 3500}{1000} = -3$	$4 \times (-3) = -12$	$(-3)^2 = 9$	$4 \times 9 = 36$
1500	6	$\frac{1500 - 3500}{1000} = -2$	$6 \times (-2) = -12$	$(-2)^2 = 4$	$6 \times 4 = 24$
2500	10	$\frac{2500 - 3500}{1000} = -1$	$10 \times (-1) = -10$	$(-1)^2 = 1$	$10 \times 1 = 10$
3500	30	$\frac{3500 - 3500}{1000} = 0$	$30 \times 0 = 0$	$(0)^2 = 0$	$30 \times 0 = 0$
4500	15	$\frac{4500 - 3500}{1000} = 1$	$15 \times 1 = 15$	$(1)^2 = 1$	$15 \times 1 = 15$
5500	10	$\frac{5500 - 3500}{1000} = 2$	$10 \times 2 = 20$	$(2)^2 = 4$	$10 \times 4 = 40$
Total	75		1		125

$$N = \text{Total frequency}$$

$$= 75$$

$$\sum fd = 1$$

$$\sum fd^2 = 125$$

$$\sigma = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} \times c$$

$$= \sqrt{\frac{125}{75} - \left(\frac{1}{75}\right)^2} \times 1000$$

$$= \sqrt{\frac{125}{75} - \frac{1}{75^2}} \times 1000$$

$$= \sqrt{\frac{9375 - 1}{75^2}} \times 1000$$

$$= \frac{\sqrt{9374}}{75} \times 1000$$

$$= \frac{\sqrt{9374}}{3} \times 40$$

value of $\sqrt{9374}$ is got by using log tables

$$\log 9374 = 3.9719$$

$$1/2 \times \log 9374 = \frac{3.9719}{2} = 1.9859$$

$$\text{Anti-log of } 1.9859 = 96.81$$

$$\sqrt{9374} = 96.81$$

$$\therefore \sigma = \frac{96.81}{3} \times 40$$

$$= 32.27 \times 40 = 1290.80$$

\therefore Standard deviation = Rs. 1290.80

Example 12:

Find the standard deviation for the following frequency distribution.

Heights in inches	No. of students with this height
59 – 61	3
61 – 63	12
63 – 65	54
65 – 67	111
67 – 69	128
69 – 71	85
71 – 73	30
73 – 75	6
75 – 77	1

Let A = midvalue of the fifth class = 68

$$c = 2$$

Space for hints

Midvalue	Frequency	$\frac{x - A}{c} = d$	fd	d ²	fd ²
x	f				
60	3	-4	-12	16	48
62	12	-3	-36	9	108
64	54	-2	-108	4	216
66	111	-1	-111	1	111
68	128	0	0	0	0
70	85	1	85	1	85
72	30	2	60	4	120
74	6	3	18	9	54
76	1	4	4	16	16
Total	430		-100		758

$N = \Sigma f = 430$

$\Sigma fd = -100$

$\Sigma fd^2 = 758$

$\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \times c$

$= \sqrt{\frac{758}{430} - \left(\frac{-100}{430}\right)^2} \times 2 \text{ inches}$

$= \sqrt{\frac{758}{430} - \left(\frac{100 \times 100}{430 \times 430}\right)} \times 2 \text{ inches}$

$= \sqrt{\frac{75.8}{43} - \frac{100}{43^2}} \times 2 \text{ inches}$

$= \sqrt{\frac{3259.4 - 100}{43^2}} \times 2 \text{ inches}$

$= \sqrt{\frac{3159.4}{43^2}} \times 2 \text{ inches}$

$= \frac{\sqrt{3159.4}}{43} \times 2 \text{ inches}$

$\log 3159.4 = 3.4995$

$$\frac{3.4995}{2} = 1.7498$$

$$\text{Anti-log of } 1.7498 = 56.20$$

$$\therefore \sqrt{3159.4} = 56.2$$

$$\therefore \sigma = \sqrt{3159.4} \times \frac{2}{43} = 56.2 \times \frac{2}{43}$$

$$= \frac{112.4}{43} = 2.6 \text{ inches}$$

$$\text{Standard deviation} = 2.6 \text{ inches}$$

Example 13:

Compute standard deviation from the following table.

Age	20-25	25-30	30-35	35-40	40-45	45-50	50-55	55-60	60-65	65-70	70-75
No.	33	112	152	154	136	118	96	74	54	37	34

Let A = midvalue of the sixth class

$$= 47.5$$

$$c = 5$$

Class	Midvalue x	Frequency f	$\frac{x - A}{c} = d$	fd	d ²	fd ²
20-25	22.5	33	-5	-165	25	825
25-30	27.5	112	-4	-448	16	1792
30-35	32.5	152	-3	-456	9	1368
35-40	37.5	154	-2	-308	4	616
40-45	42.5	136	-1	-136	1	136
45-50	47.5	118	0	0	0	0
50-55	52.5	96	1	96	1	96
55-60	57.5	74	2	148	4	296
60-65	62.5	54	3	162	9	486
65-70	67.5	37	4	148	16	592
70-75	72.5	34	5	170	25	850
Total		1000		-789		7057

Space for hints

Space for hints

$$N = \Sigma f = 1000$$

$$\Sigma fd = -789$$

$$\Sigma fd^2 = 7057$$

$$\therefore \sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \times c$$

$$= \sqrt{\frac{7057}{1000} - \left(\frac{-789}{1000}\right)^2} \times 5$$

$$= \sqrt{7.057 - .6226} \times 5$$

$$= \sqrt{6.4344} \times 5$$

$$= 2.536 \times 5 = 12.680$$

$$\therefore \text{Standard deviation} = 12.68$$

Example 14:

Calculate the standard deviation.

	f
4000 to 5000	16
3000 „ 4000	48
2000 „ 3000	98
1000 „ 2000	121
0 „ 1000	62
-1000 „ 0	40
-2000 „ -1000	30
-3000 „ -2000	16
-4000 „ -3000	4

Let A = midvalue of the fifth class.

$$\therefore A = 500$$

$$c = 1000$$

Space for hints

Midvalue x	Frequency f	$\frac{x - A}{c} = d$	fd	d ²	fd ²
4500	16	4	64	16	256
3500	48	3	144	9	432
2500	98	2	196	4	392
1500	121	1	121	1	121
500	62	0	0	0	0
-500	40	-1	-40	1	40
-1500	30	-2	-60	4	120
-2500	16	-3	-48	9	144
-3500	4	-4	-16	16	64
Total	435		361		1569

$$N = \Sigma f = 435$$

$$\Sigma fd = 361$$

$$\Sigma fd^2 = 1569$$

$$\therefore \sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \times c$$

$$= \sqrt{\frac{1569}{435} - \left(\frac{361}{435}\right)^2} \times 1000$$

$$= \sqrt{\frac{1569 \times 435 - (361)^2}{435^2}} \times 1000$$

$$= \sqrt{\frac{682515 - 130321}{435^2}} \times 1000$$

$$= \sqrt{\frac{552194}{435^2}} \times 1000$$

$$= 1.708 \times 1000$$

$$= 1708$$

$$\therefore \text{Standard deviation} = 1708$$

8.9 Properties of Standard Deviation :

1. For a given distribution, the sum of squares of the deviations of various items from mean is always less than the sum of the squares of the deviations from median or mode. Therefore, the arithmetic mean of the squares of the deviations from mean is always less than the arithmetic mean of the squares of the deviations from median or mode. Because of this special characteristic of mean, standard deviation gives the amount of the dispersion more accurately.

2. Suppose \bar{x}_1 and σ_1 are the mean and standard deviation of one group of n_1 items.

Suppose \bar{x}_2 and σ_2 are the mean and standard deviation of another group of n_2 items.

Now the standard deviation of the combination of the two groups of items is given by the following formula,

$$\sigma = \sqrt{\frac{1}{N} (n_1 \sigma_1^2 + n_2 \sigma_2^2 + n_1 d_1^2 + n_2 d_2^2)}$$

Where σ denotes the standard deviation of the combined group of items.

N denotes the total number of items in the combined group (i.e.) $N = (n_1 + n_2)$

d_1 denotes the value of $(\bar{x}_1 - \bar{x})$

d_2 denotes the value of $(\bar{x}_2 - \bar{x})$

\bar{x} denotes the mean value of the combined group which is calculated using the formula

$$\bar{x} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2}$$

The above formula is given in another form as follows:

$$\sigma = \sqrt{\frac{1}{N} \left(n_1 \sigma_1^2 + n_2 \sigma_2^2 + \frac{n_1 n_2}{N} (\bar{x}_1 - \bar{x}_2)^2 \right)}$$

Example 15:

Find out the combined standard deviation from the following data.

	Group A	Group B
Number of items	100	500
Mean	50	60
Standard deviation	10	11

Check your Progress

17. What are the important properties of S.D.?

$$n_1 = \text{number of items in A} = 100$$

$$n_2 = \text{number of items in B} = 500$$

$$\bar{x}_1 = \text{Mean of items in A} = 50$$

$$\bar{x}_2 = \text{Mean of items in B} = 60$$

$$\sigma_1 = \text{standard deviation of items in A} = 10$$

$$\sigma_2 = \text{standard deviation of items in B} = 11$$

$$\bar{x} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2}$$

$$= \frac{(100 \times 50) + (500 \times 60)}{100 + 500}$$

$$= \frac{5000 + 30000}{600}$$

$$= \frac{35000}{600} = \frac{350}{6} = \frac{175}{3}$$

$$(n_1 + n_2) = 100 + 500$$

$$= 600$$

$$\therefore N = 600$$

$$d_1 = \bar{x}_1 - \bar{x}$$

$$= 50 - \frac{175}{3}$$

$$= \frac{150 - 175}{3}$$

$$= \frac{-25}{3}$$

$$d_2 = \bar{x}_2 - \bar{x}$$

$$= 60 - \frac{175}{3}$$

Space for hints

$$= \frac{180 - 175}{3} = \frac{5}{3}$$

$$\sigma = \sqrt{\frac{1}{N} \left[(n_1 \sigma_1^2 + n_2 \sigma_2^2 + n_1 d_1^2 + n_2 d_2^2) \right]}$$

$$\sigma = \sqrt{\frac{1}{600} \left[100 \times (10)^2 + 500 \times (11)^2 + 100 \times \left[\frac{-25}{3} \right]^2 + 500 \times \left[\frac{5}{3} \right]^2 \right]}$$

$$\sigma = \sqrt{\frac{1}{600} \left[(100 \times 100) + (500 \times 121) + \left(100 \times \frac{625}{9} \right) + \left(500 \times \frac{25}{9} \right) \right]}$$

$$= \sqrt{\frac{1}{600} \left[10000 + 60500 + \frac{62500 + 12500}{9} \right]}$$

$$= \sqrt{\frac{1}{600} \left[\frac{90000 + 544500 + 62500 + 12500}{9} \right]}$$

$$= \sqrt{\frac{1}{600} \times \frac{709500}{9}}$$

$$= \sqrt{\frac{7095}{6 \times 9}}$$

$$= \sqrt{\frac{2365}{2 \times 9}}$$

$$= \sqrt{\frac{1182.5}{9}}$$

$$= \frac{\sqrt{1182.5}}{3}$$

$\sqrt{1182.5}$ is found out using log table as follows:

$$\log 1182.5 = 3.0727$$

$$\frac{1}{2} \times \log 1182.5 = \frac{3.0727}{2} = 1.5363$$

$$\text{Anti-log } 1.5363 = 34.38$$

$$\sqrt{1182.5} = 34.38$$

$$\therefore \sigma = \sqrt{\frac{1182.5}{3}} = \frac{34.38}{3} = 11.46$$

\therefore Combined standard deviation = 11.46

We can calculate the standard deviation of the combined set using the other formula as follows.

$$\begin{aligned}\sigma &= \sqrt{\frac{1}{N} \left(n_1\sigma_1^2 + n_2\sigma_2^2 + \frac{n_1n_2}{N} + (\bar{x}_1 - \bar{x}_2)^2 \right)} \\ \sigma &= \sqrt{\frac{1}{600} \left[100 \times (10)^2 + 500 \times (11)^2 + \frac{100 \times 500}{600} (50 - 60)^2 \right]} \\ &= \sqrt{\frac{1}{600} \left(10,000 + 500 \times 121 + \frac{500}{6} (-10)^2 \right)} \\ &= \sqrt{\frac{1}{600} \left(10,000 + 60,500 + \frac{250}{3} \times 100 \right)} \\ &= \sqrt{\frac{1}{600} \left(\frac{30,000 + 181,500 + 25,000}{3} \right)} \\ &= \sqrt{\frac{1}{600} \times \frac{2,36,500}{3}} \\ &= \sqrt{\frac{2365}{6 \times 3}} \\ &= \sqrt{\frac{1182.5}{3 \times 3}}\end{aligned}$$

We have already found out the value of $\sqrt{\frac{1182.5}{3 \times 3}}$ to be 11.46

$$\therefore \sigma = 11.46$$

\therefore Combined standard deviation = 11.46.

It is to be noted that the values of combined standard deviation obtained with the help of both the formulae are the same. Therefore, the student can use any one of the two formulae to get the combined standard deviation.

Example 16:

The following results were found in an investigation

	Factory A	Factory B
No. of workers	300	200
Average of wages in Rs.	52	40
Variance of wages in Rs.	81	64

Calculate the combined standard deviation of wages in both factories together.

$$n_1 = 300 \quad n_2 = 200$$

$$\bar{x}_1 = 52 \quad \bar{x}_2 = 40$$

$$\sigma_1^2 = 81 \quad \sigma_2^2 = 64$$

$$\therefore N = n_1 + n_2$$

$$= 300 + 200 = 500$$

$$\sigma = \sqrt{\frac{1}{N} \left(n_1 \sigma_1^2 + n_2 \sigma_2^2 + \frac{n_1 n_2}{N} (\bar{x}_1 - \bar{x}_2)^2 \right)}$$

$$= \sqrt{\frac{1}{500} \left[(300 \times 81) + (200 \times 64) + \left(\frac{300 \times 200}{500} (52 - 40)^2 \right) \right]}$$

$$= \sqrt{\frac{1}{500} [(24300 + 12800 + 120 \times 144)]}$$

$$= \sqrt{\frac{1}{500} (24300 + 12800 + 17280)}$$

$$= \sqrt{\frac{54380}{500}}$$

$$= \sqrt{108.76}$$

$$= 10.43$$

\therefore Combined standard deviation = Rs. 10.43.

8.10 COEFFICIENT OF VARIATION (C.V.)

Space for hints

We have seen a measure of dispersion called standard deviation, in the previous topic. It is an absolute measure of dispersion. An absolute measure of dispersion is always expressed in terms of the same units in which the given distribution is expressed. Therefore standard deviation is expressed in the units in which the given distribution is expressed. Because of this reason, the dispersions of two distributions given in two different units cannot be compared with the help of their standard deviations. Hence, it becomes necessary to calculate a relative measure of dispersion to facilitate comparison between two distributions given in two different units.

The relative measure of dispersion corresponding to standard deviation is called "coefficient of standard deviation". The formula to calculate the coefficient of standard deviation is given as follows:

$$\text{Coefficient of standard deviation} = \frac{\text{Standard deviation}}{\text{Mean}} = \frac{\sigma}{\bar{x}}$$

Coefficient of standard deviation is usually expressed in percentage units and it is called coefficient of variation (C.V.). Therefore, the formula to calculate the coefficient of variation is as follows:

$$\text{Coefficient of variation (C.V.)} = \text{Coefficient of standard deviation} \times 100$$

$$= \frac{\sigma}{\bar{x}} \times 100$$

Coefficient of variation has much practical utility. When two groups of data are given the group of data for which coefficient of variation is less is considered less variable, more stable, more uniform or more consistent.

On the other hand, that group of data for which coefficient of variation is greater will be more variable, less stable, less uniform or less consistent.

Example 17:

The means and standard deviations of the marks obtained by two students X and Y are given below. Which of the two candidates is more consistent in his performance?

	X	Y
Mean	72	78
Standard deviation	8	6

Check your Progress

18. What is C.V.? Give the formula.

Coefficient of variation of X:

Mean of the marks obtained by X = 72

Standard deviation of the marks obtained by X = 8

$$\% \text{ Coefficient of Variation of } x = \frac{8}{72} \times 100 = 11.11\%$$

Coefficient of variation of Y:

Mean of the marks obtained by Y = 78

Standard deviation of the marks obtained by Y = 6

$$\% \text{ Coefficient of Variation of } Y = \frac{6}{78} \times 100 = 7.7\%$$

Coefficient of variation of Y is less than the coefficient of variation of X.

\therefore Y is more consistent than X.

That is, the candidate Y is more consistent in his performance than the candidate X.

Example 18 :

The following table gives the means and standard deviations of runs scored by two batsmen A and B.

Who is the more consistent batsman?

	Mean	Standard deviation
A	39	32
B	43	35

$$\text{Coefficient of variation of batsman A} = \frac{\text{Standard deviation}}{\text{Mean}} \times 100$$

$$= \frac{32}{39} \times 100 = \frac{3200}{39} = 82.05\%$$

$$\text{Coefficient of variation of batsman B} = \frac{\text{Standard deviation}}{\text{Mean}} \times 100$$

$$= \frac{35}{43} \times 100 = \frac{3500}{43} = 81.4\%$$

Coefficient of variation of runs for B viz., 81.4% is less than that for A viz., 82.05%

Space for hints

∴ Batsman B is more consistent in his batting.

Example 19 :

Two cricketers scored the following runs in the several innings. Find who is a better run getter and who is more consistent player.

A:	42	17	83	59	72	76	64	65	45	40	32
B:	28	70	31	0	59	108	82	14	3	95	64

Arithmetic mean of the runs scored by a cricketer will give us an idea whether he is a better run getter or not. The cricketer having a greater arithmetic mean of the runs scored is considered to be the better run getter. So, for the two cricketers first we calculate the arithmetic means of the runs scored by them.

Mean for A

$$\begin{aligned}\bar{x} &= \frac{\sum x}{N} \\ &= \frac{42 + 17 + 83 + 59 + 72 + 76 + 64 + 65 + 45 + 40 + 32}{11} \\ &= \frac{595}{11} = 54.09 = 54 \text{ (approximately)}\end{aligned}$$

Mean for B

$$\begin{aligned}\bar{x} &= \frac{\sum x}{N} \\ &= \frac{28 + 70 + 31 + 0 + 59 + 108 + 82 + 14 + 3 + 95 + 64}{11} \\ &= \frac{554}{11} = 50.3 = 50 \text{ (approximately)}\end{aligned}$$

Arithmetic mean of the runs scored by A viz., 54 is greater than the arithmetic mean of the runs scored by B, viz., 50.

A is the better run getter.

To find out who is the more consistent player we have to calculate the coefficient of variation of the runs scored by each cricketer. For this, first we calculate the standard deviation for each player as follows:

Space for hints

S.D. for A

Let the origin, A = 55.

x	(x-A) = d	d ²
42	-13	169
17	-38	1444
83	28	784
59	4	16
72	17	289
76	21	441
64	9	81
65	10	100
45	-10	100
40	-15	225
32	-23	529
Total	-10	4178

$$\Sigma d = -10,$$

$$\Sigma d^2 = 4178$$

$$\sigma = \sqrt{\frac{\Sigma d^2}{n} - \left(\frac{\Sigma d}{n}\right)^2}$$

$$= \sqrt{\frac{4178}{11} - \left(\frac{-10}{11}\right)^2}$$

$$= \sqrt{\frac{4178}{11} - \frac{100}{121}}$$

$$= \sqrt{\frac{45958 - 100}{121}}$$

$$= \sqrt{\frac{45858}{121}}$$

$$= \frac{214.2}{11}$$

$$= 19.47$$

S.D. for B

Let the origin, A = 50

Space for hints

x	(x-A) = d	d ²
28	-22	484
70	20	400
31	-19	361
0	-50	2500
59	9	81
108	58	3364
82	32	1024
14	-36	1296
3	-47	2209
95	45	2025
64	14	196
Total	4	13940

$$\Sigma d = 4$$

$$\Sigma d^2 = 13940$$

$$\sigma = \sqrt{\frac{\Sigma d^2}{n} - \left(\frac{\Sigma d}{n}\right)^2}$$

$$= \sqrt{\frac{13940}{11} - \left(\frac{4}{11}\right)^2}$$

$$= \sqrt{\frac{13940}{11} - \frac{16}{121}}$$

$$= \sqrt{\frac{153340 - 16}{121}}$$

$$= \frac{\sqrt{153324}}{11}$$

$$= \frac{391.4}{11}$$

$$= 35.6 \text{ (approx.)}$$

Now let us calculate the coefficient of variation for A and B as follows.

C.V. for A:

σ	=	19.47
\bar{x}	=	54
$\% \text{ c.v.}$	=	$\frac{\sigma}{\bar{x}} \times 100$
	=	$\frac{19.47}{54} \times 100 = 36.05\%$

C.V. for B:

$$\begin{aligned} \% \text{ c.v.} &= \frac{\sigma}{\bar{x}} \times 100 \\ &= \frac{35.6}{50} \times 100 = 71.2\% \end{aligned}$$

C.V. for A is less than C.V. for B. Therefore, A is the more consistent player.

Ans: A is the better run getter as well as more consistent player.

Example 20:

The following results were found in an investigation.

	Factory A	Factory B
No. of workers	300	200
Average wages in Rs.	52	40
Variance of wages in Rs.	81	64

State in which factory wages are more variable?

For Factory A:

Space for hints

$$\bar{x} = 52$$

$$\sigma^2 = 81$$

$$\therefore \sigma = \sqrt{81} = 9$$

$$\text{Co-efficient of variation} = \frac{\sigma}{\bar{x}} \times 100$$

$$= \frac{9}{52} \times 100$$
$$= 17\%$$

For Factory B:

$$\bar{x} = 40$$

$$\sigma^2 = 64$$

$$\therefore \sigma = \sqrt{64} = 8$$

$$\text{Co-efficient of variation} = \frac{\sigma}{\bar{x}} \times 100$$

$$= \frac{8}{40} \times 100$$
$$= 20\%$$

Co-efficient of variation for Factory B is greater than that for Factory A.

\therefore Wages are more variable in Factory B

8.11 RELATIVE ADVANTAGES AND DISADVANTAGES OF VARIOUS MEASURES OF DISPERSION

I. Properties of a good measure of dispersion.

A good measure of dispersion should possess most of the following qualities.

- (1) It should be based on the values of all the items.
- (2) It should be easily calculated.
- (3) It should be readily understood.
- (4) It should be least affected by fluctuations of sampling and
- (5) It should lend itself for further algebraic treatment.

Absolute Measures of Dispersion

II. Relative merits and demerits of Absolute Measures of Dispersion :

(A) RANGE

(a) Merits :

- (1) Range is easily calculated.
- (2) It is easily understood.

(b) Demerits :

- (1) It is not based on each and every item of the distribution.
- (2) It is easily affected by fluctuations of sampling.
- (3) It does not give any idea about the item between the two extremes.
- (4) It cannot be computed in the case of open end distribution.
- (5) It fails to consider the central tendency of the data

(c) Uses :

- (1) It is useful in studying the variations in the prices of stocks and shares and other commodities. By computing the range, we get an idea about the variation in the price level.
- (2) The information about the minimum and maximum temperature and their difference is essential for the public.
- (3) A doctor will be keen on the range of the varying temperatures of his patients.

(B) QUARTILE DEVIATION (Q.D.)

Space for hints

(a) Merits :

- (1) It is commonly understood by all.
- (2) It is easily calculated and can also be graphically calculated using Ogive.
- (3) Though it is not based on all the values of the items, all items play an indirect role in its determination
- (4) The extreme values do not affect this.
- (5) It can be calculated in case of open end distribution, an advantage over "Range".

(b) Demerits :

- (1) It is not capable of further algebraic treatment.
- (2) It ignores the first and last 25% of the items.
- (3) It is affected to a considerable extent by the fluctuations of sampling.

(c) Uses :

Though it is used only as a rough measure, it is used in calculating the skewness and kurtosis of a frequency distribution.

(C) MEAN DEVIATION (M.D.)

(a) Merits:

- (1) Unlike Q.D. or range, it is a measure based on the values of all the items.
- (2) It is readily understood. Anyone who is familiar with the concept of average can easily understand the meaning of the mean deviation.

(b) Demerits :

- (1) It is not suitable for algebraic treatment, as we have ignored the signs of the deviations.
- (2) It is not as easily understood by the layman as the quartile deviation.
- (3) It is less stable than the S.D.

(c) Uses :

It is mostly used only in economic statistics. It is greatly used in situations where the degree of dispersion of a distribution is to be shown to the general public who are not well grounded in statistics.

The National Bureau of Economic Research of the U.S.A. has found it to be the most practical measure of dispersion that can be used in its work of forecasting business cycles.

(D) STANDARD DEVIATION (S.D.)

It is considered to be the best measure, as it satisfies most of the properties required by a good measure of dispersion.

(a) Merits :

- (1) It is based on all the items.
- (2) It is rigidly defined.
- (3) It lends itself for further algebraic treatment
 - (i) the S.D. of the combined distribution of two or more distributions can be calculated.
 - (ii) It possesses many mathematical properties and hence it is used in advanced studies.
- (4) It is not affected considerably by the fluctuations of sampling

(b) Demerits :

- (1) It is not easy to understand by the ordinary layman.
- (2) It is difficult to calculate (Compared to other measures).
- (3) It gives more weight to extreme items and less weight to items which are nearer to mean. For, the squares of the deviation which are big in size, would be proportionately greater than the squares of the deviations which are comparatively small.

For example, suppose the deviations are 2 and 8. These deviations are in the ratio 1 : 4. Squares of these deviations are ($4 = 2^2$) and ($64 = 8^2$) and their ratio is 1 : 16.

(c) Uses :

Space for hints

- (1) It is used in further statistical work-for example, in computing skewness, correlation etc., use is made of standard deviation.
- (2) It is of greatest value in testing the reliability of measures computed from samples.
- (3) It is considered to be the best measures of dispersion and hence it is very commonly used.

III. Choice of a Measure of Dispersion

The choice of a measure of dispersion depends on the nature, purpose and object of investigation. Let us take the four measures one by one and see whether they are preferable.

Range can be used only as a rough and ready measure of spread of data. It can also be used when the data set is incomplete or when the variation in the size of the items is very small.

Quartile Deviation is a better measure than the range as it is not affected by the values of extreme items. It can be used only in those cases where mean deviation or standard deviation cannot be easily calculated.

Mean deviation has advantage over standard deviation as it can be easily calculated and easily understood. Also, mean deviation is minimum when calculated from the median. So, where median is supposed to be the best average, the suitable measure of dispersion would be the mean deviation. The economists and businessmen prefer this measure because of the ease in calculation.

Standard deviation is the most useful of all the measures of dispersion as it possesses most of the properties required by a good measure. It is the most stable measure differing very little from sample to sample. It occupies the same distinct position among the measures of dispersion that the A.M. does among the averages. Sum of the squares of the deviations of all the items from the mean is minimum and hence, the standard deviation is the minimum root mean square deviation. For all practical purposes S.D. should be preferred because of its greater accuracy and precision.

Check your Progress

19. Which is the best measure of dispersion? Why?

To conclude, it can be stated that if statistical reliability is the prime factor to be considered, the order of preference is S.D., M.D., Q.D., and the range. But if ease and quickness alone are to be considered, the order is just the reverse.

Let us also note the empirical relation between the three measures.

$$Q.D., = \frac{2}{3} S.D.,$$

$$M.D., = \frac{4}{5} S.D.$$

IV. Advantages of Relative Measures of Dispersion

(a) Avoiding misleading conclusion:

For the comparison of two distributions, we should always choose a relative measure of dispersion. Absolute measures of dispersion sometimes, gives us very misleading conclusions. For example, suppose the profits of two Companies A and B during the last three years are as follows:

A (Rs.)	B (Rs.)
1250	3250
2000	4000
2750	4750

Ranges of the profits of both A and B are the same. If is equal to Rs. 1500.

Mean deviation in both the cases equal to Rs. 500.

Standard deviation in both the cases equal to Rs. 612.4.

Thus the absolute measure of dispersion of the profits of the two companies are equal. Hence, we may derive the conclusion that the

dispersions of the profits of the two companies are equal. But we can prove that this conclusion is a fallacious one by the following :

Space for hints

$$\text{Coefficient of range for A} = \frac{3}{8}$$

$$\text{Coefficient of range for B} = \frac{3}{16}$$

∴ Coefficient of range for B is less than that for A.

$$\text{Coefficient of mean deviation for A} = \frac{1}{4}$$

$$\text{Coefficient of mean deviation for B} = \frac{1}{8}$$

∴ Coefficient of mean deviation for B is less than that for A.

$$\text{Coefficient of variation for A} = 30.62\%$$

$$\text{Coefficient of variation for B} = 15.31\%$$

Coefficient of variation for B is less than that for A.

Thus, all the relative measures of dispersion for B are less than that for A. From this it follows that B's profits have a lesser variability than those of A's. That is, B's profits are more stable than those of A's. Coefficient of dispersion, therefore, corrects the wrong impression created by the absolute measures.

(b) Facilitating comparisons of two or more distributions given in different units :

In comparing dispersion of two distributions expressed in two different units, the use of relative measures of dispersion is inevitable.

For, an absolute measure of dispersion is expressed in the same unit in which the given distribution is expressed and a relative measure is free from units of measurement and is a pure number.

Here again, we consider coefficient of variation $\left(\frac{\sigma}{\bar{x}} \times 100\right)$ as the appropriate one because of the importance of A.M. and S.D.

II. MEASURES OF SKEWNESS

1. Skewness - Meaning

Averages and measures of dispersion tell us about the two important aspects Viz., the central value and the concentration of items around the central value of a frequency distribution. The third important aspect of a frequency distribution is its "skewness". Skewness means lacking in symmetry.

2. Symmetrical Distribution - Meaning :

A distribution is said to be symmetrical when the frequencies are symmetrically distributed about the mode : that is, when the values equidistant from the mode have equal frequencies. The following is an example of symmetrical distribution.

Value of item	Frequency
1	3
2	4
3	6
4	9
5	10
6	9
7	6
8	4
9	3

Mode of the above distribution is 5 ; 4 and 6 are the values equidistant from the modal value, and they have the same frequency viz., 9. Similarly, 3 and 7 are values "equidistant from the modal value 5 and have the same frequency viz., 6. In the same way, the pairs of values 2 and 8, 1 and 9 are equidistant from the modal value and have the same frequencies viz. 4 and 3 respectively.

Any distribution which is of the above form is called a symmetrical distribution. A symmetrical distribution has the following properties:

1. In a symmetrical distribution the values of mean and median coincide with the value of mode.

That is, $a = M = Z$

where a denotes mean; M denotes median; Z denotes mode.

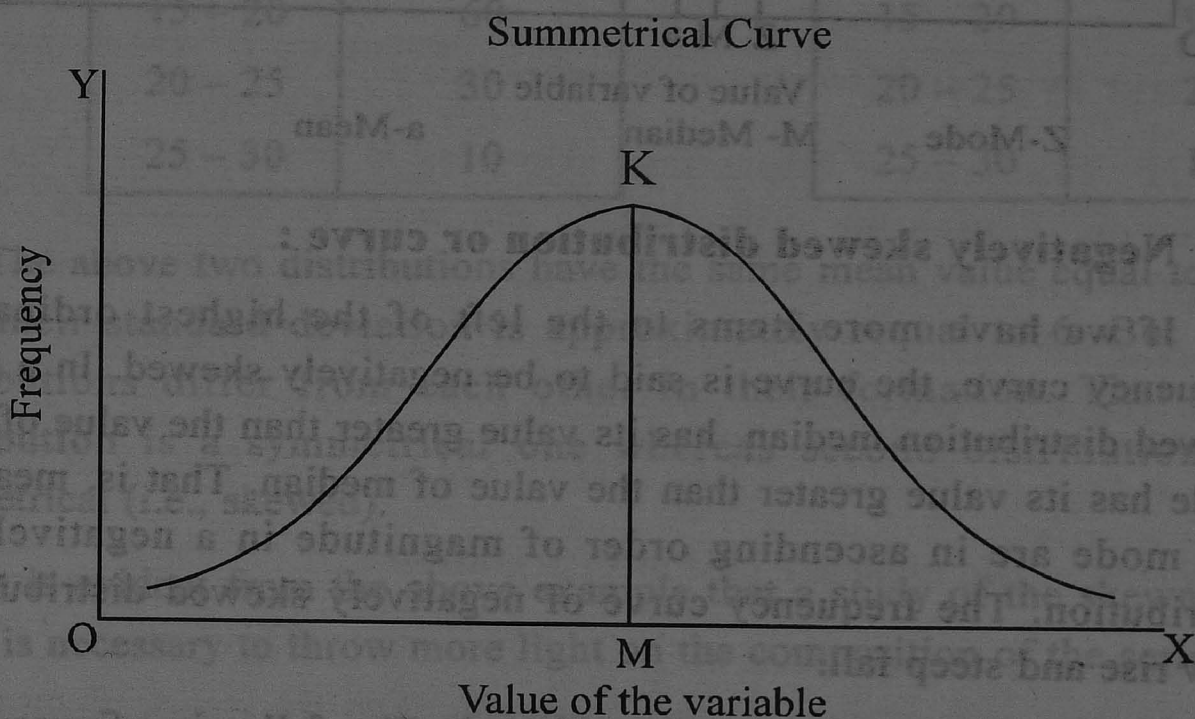
2. Also in this case, median lies half way between the lower and upper quartiles,

Space for hints

$$\text{That is, } (Q_3 - M) = (M - Q_1)$$

where Q_3 denotes upper quartile; M denotes median; Q_1 denotes lower quartile.

3. In a symmetrical distribution, the sum of the positive deviations from median will be equal to the sum of the negative deviations from median.
4. If we construct the frequency curve of a symmetrical distribution, it will be a perfectly bellshaped curve. The following figure shows the shape of the frequency curve of a symmetrical distribution.



When the given distribution is not a symmetrical distribution we say that the distribution is skewed; Its frequency curve is said to be skewed.

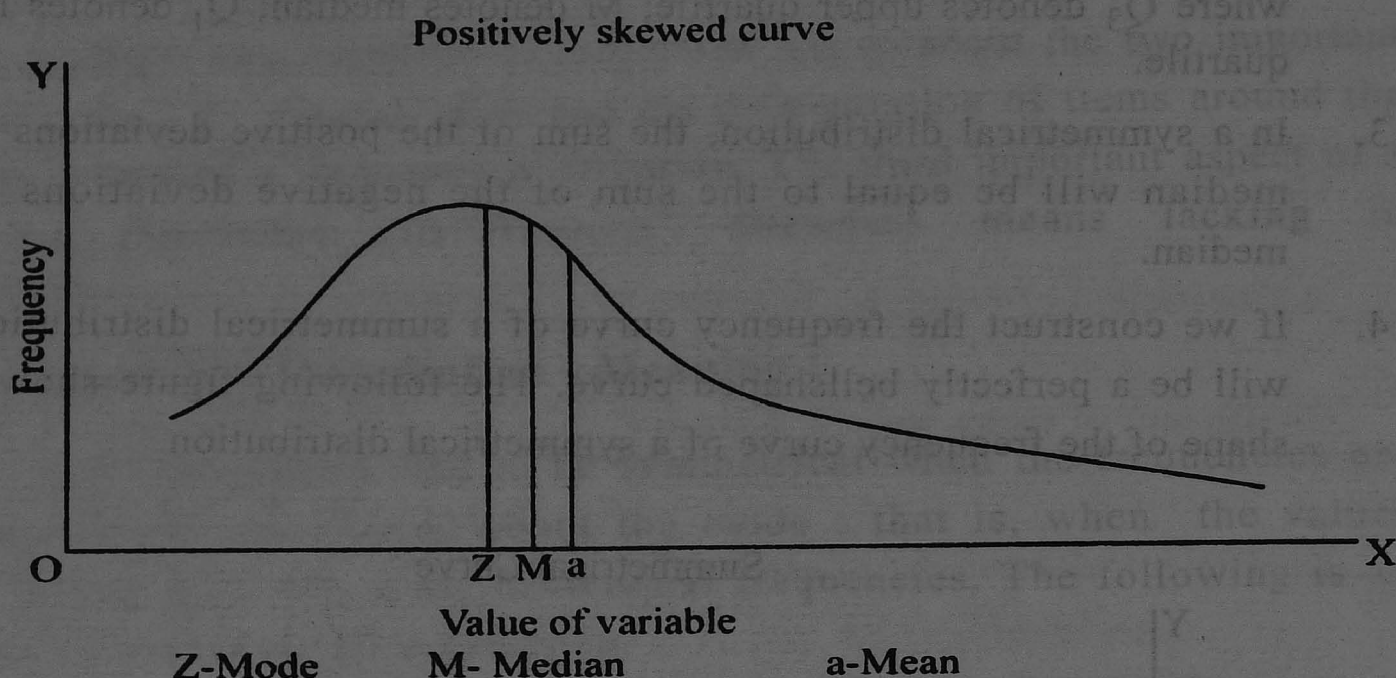
3. Types of Skewness :

Skewed distributions are of two types (i) positively skewed and (ii) negatively skewed.

(i) Positively skewed distribution or curve :

If we have more items to the right of the highest ordinate of the frequency curve, the curve is said to be positively skewed. In a positively skewed distribution median has its value greater than the value of mode and mean has its value greater than the value of the median. That is, mode, median and mean are in ascending order of magnitude in a positively skewed distribution. The frequency curve of a positively skewed distribution has a steep rise and slow fall (i.e.,) it has a long tail at the right. The following

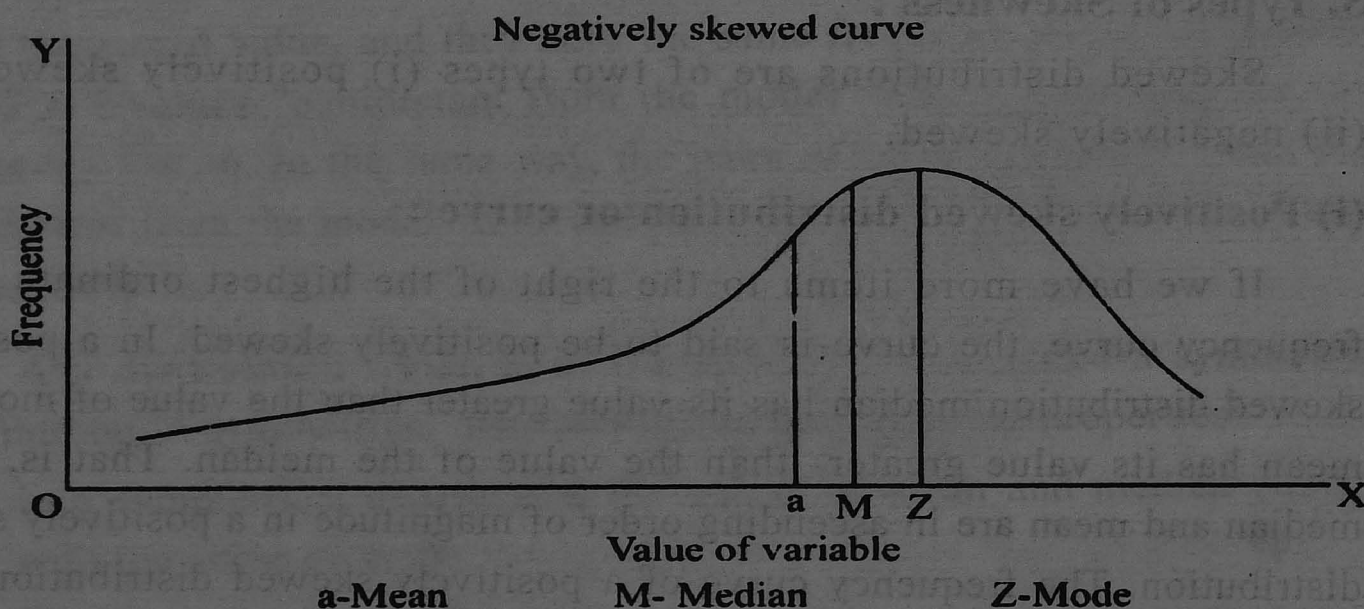
figure shows the shape of the frequency curve of a positively skewed distribution.



(ii) Negatively skewed distribution or curve :

If we have more items to the left of the highest ordinate of the frequency curve, the curve is said to be negatively skewed. In a negatively skewed distribution median has its value greater than the value of mean and mode has its value greater than the value of median. That is, mean, median and mode are in ascending order of magnitude in a negatively skewed distribution. The frequency curve of negatively skewed distribution has a slow rise and steep fall.

That is, it has a long tail at the left. The following figure shows the shape of the frequency curve of a negatively skewed distribution.



4. Need for a measure of Skewness :

Space for hints

Two distributions may have the same mean and standard deviation and yet they differ from each other in their formation. Consider the following two frequency distributions.

Distribution I

Class	Frequency
0 – 5	10
5 – 10	30
10 – 15	60
15 – 20	60
20 – 25	30
25 – 30	10

Distribution II

Class	Frequency
0 – 5	10
5 – 10	40
10 – 15	30
15 – 20	90
20 – 25	20
25 – 30	10

The above two distributions have the same mean value equal to 15 and have their standard deviation is approximately requal to 6. But the two distributions differ from each other in their formation. That is, first distribution is a symmetrical one whereas second distribution is not symmetrical (i.e., skewed).

It is evident from the above example that a study of the skewness of a series is necessary to throw more light on the composition of the series.

5. Tests of Skewness :

1. In a skewed distribution value of mean, median and mode do not coincide. They satisfy the relation.

$$\text{Mean} - \text{Mode} = 3 (\text{Mean} - \text{Median})$$

$$(\text{i.e.,}) (a - Z) = 3 (a - M)$$

2. The frequency curve of a skewed distribution is not perfectly bell-shaped. It has a tail at the left or at the right.
3. For a skewed distribution, the sum of the positive deviations is not equal to the sum of the negative deviations from median.
4. In a skewed distribution, median would not be halfway between the lower and upper quartiles. That is $(Q_3 - M) - (M - Q_1)$ would not be equal to zero.
5. In a skewed distribution, the values equidistant from mode do not have equal frequencies.

6. Measures of Skewness :

With the help of the tests of skewness given above we can find out whether a particular distribution is skewed, or not. If the tests of skewness indicate that a particular distribution is skewed, the next problem that arises is to measure the extent of skewness. Some distributions may differ slightly from a symmetrical distribution while other distributions may differ widely from it. Measures of skewness are meant to give an idea about the amount of asymmetry in a series.

We have two types of measures of skewness. The first type of measure is given in terms of the measures of central tendency viz., Mean, Median and Mode. The second type of measure of skewness is given in terms of quartiles.

6.1 First Type of Measure of Skewness :

The first type of measure of skewness is based on the fact that in a skewed distribution the mean, median and mode do not coincide. The larger the difference between any two of these values, the greater is the degree of skewness and vice versa. We take the difference between mean and mode as the measure of skewness.

$$\therefore \text{Measure of skewness} = (\text{Mean} - \text{Mode}) = (a - Z)$$

6.2 Value of the measure and interpretation :

As we have stated earlier, when the given distribution is a positively skewed one, the value of mean is greater than the value of mode. In this case, the above measure of skewness will have positive sign.

Therefore, whenever the measure of skewness given above has positive sign, the distribution is said to have positive skewness.

When the given distribution is negatively skewed, the value of mean is less than the value of mode. In this case, the above measures of skewness will have negative sign.

Therefore, wherever the above measure of skewness has negative sign, the distribution is said to have negative skewness.

6.3 Relative Measure of Skewness - Karl Pearson's coefficient :

The measure given above is an absolute measure of skewness, Hence, it is expressed in the same units in which the given distribution is expressed. Therefore, it is not possible to compare the skewness of two distributions

given in two different units with the help of the above measure of skewness. For purposes of comparison, it therefore becomes necessary to have a relative measure of skewness.

Space for hints

Relative measure of skewness is got by dividing the absolute measure of skewness by a measure of dispersion. But, usually, relative measure of skewness is got by dividing absolute measure of skewness by standard deviation. The relative measure was given by Karl Pearson and hence it is called Pearson's coefficient of skewness.

$$\begin{aligned} \therefore \text{Karl Pearson's coefficient of skewness} &= \frac{\text{Mean}(a) - \text{Mode}(z)}{\text{Standard deviation}(\sigma)} \\ &= \frac{a - z}{\sigma} \end{aligned}$$

There exists an empirical relationship between Mean, Median and Mode and the relationship is as follows:

$$(\text{Mean} - \text{Mode}) = 3 (\text{Mean} - \text{Median})$$

$$\text{That is, } (a - z) = 3 (a - M)$$

Therefore, when the mode is not well-defined for the given distribution (i.e., when given distribution has more than one mode or no mode) Karl Pearson has substituted the value of $3(a - M)$ for the value of $(a - Z)$.

Hence, Pearson's coefficient of skewness is also given as follows* :

$$\text{Karl Pearson's coefficient of skewness} = \frac{3(a - M)}{\sigma}$$

6.4 Interpretation of the values of the coefficient :

The value of $\frac{3(a - M)}{\sigma}$ usually lies between -1 and 1 .

Therefore, the value of $\frac{3(a - M)}{\sigma}$ lies between -3 and 3 . Thus the value of the coefficient of skewness is either positive, zero or negative.

When the coefficient of skewness is positive, the distribution is said to have positive skewness.

When the coefficient of skewness is zero, the given distribution is symmetrical.

When the coefficient of skewness is negative, the given distribution is said to have negative skewness.

* In doing problems, we always use only this second form of Pearson's coefficient of skewness viz., $\frac{3(a - M)}{\sigma}$

Example 1 :

Find out the nature of skewness of the following data. Also find out Pearson's coefficient of skewness.

Wages (in Rs.)	Frequency
0 – 5	10
5 – 10	40
10 – 15	30
15 – 20	90
20 – 25	20
25 – 30	10

We calculate the mean as follows :

Mid value x	Frequency f	xf
2.5	10	25
7.5	40	300
12.5	30	375
17.5	90	1575
22.5	20	450
27.5	10	275
Total	200	3000

$$\Sigma f = 200 \quad \Sigma xf = 3000$$

$$\bar{x} = \frac{\Sigma xf}{\Sigma f} = \text{Rs. } \frac{3000}{200} = \text{Rs. } 15.$$

$$\text{Mean (a)} = \text{Rs. } 15$$

Median is calculated as follows :

$$N = \text{Total frequency} = 200$$

$$\frac{N}{2} = \frac{200}{2} = 100$$

$$\begin{aligned} \text{Median} &= \text{Value of the item } \left(\frac{N}{2} \right) \\ &= \text{Value of the item (100)} \end{aligned}$$

Class	Frequency	Cumulative Frequency
0 - 5	10	10
5 - 10	40	50
10 - 15	30	80
15 - 20	90	170
20 - 25	20	190
25 - 30	10	200

Space for hints

From the cumulative frequency column of the above table we come to know that all the items after the items 80 and upto the item 170 are having their values lying in the interval "15-20".

The item 100 is in between the items 80 and 170.

∴ The value of the item 100 also lies in the interval "15-20" (i.e.,) the value of median lies in the interval "15-20"

∴ "15-20" is the median class.

$$\therefore l = 15$$

$$m = 80$$

$$f = 90$$

$$c = 20 - 15 = 5$$

$$\therefore \text{Median} = l + \frac{\frac{N}{2} - m}{f} \times c$$

$$= \text{Rs.} \left[15 + \frac{100 - 80}{90} \times 5 \right]$$

$$= \text{Rs.} \left[15 + \frac{20}{18} \right]$$

$$= \text{Rs.} \left[15 + \frac{10}{9} \right]$$

$$= \text{Rs.} [15 + 1.11]$$

$$= \text{Rs.} 16.11$$

$$\therefore M = 16.11$$

Midvalue	Frequency				
2.5	10				
7.5	40				
12.5	30				
17.5	90				
22.5	20				
27.5	10				
Total	200				

Space for hints

Mode is calculated as follows :

Highest frequency in the given distribution = 90

The class having this highest frequency viz., 90 is 15-20

∴ Modal class is "15-20"

$$\therefore l = 15$$

$$f_1 = 30$$

$$f_2 = 20$$

$$c = (20 - 15) = 5$$

$$\text{Mode} = l + \frac{cf_2}{f_1 + f_2} = \text{Rs.} \left[15 + \frac{5 \times 20}{30 + 20} \right] = \text{Rs.} \left[15 + \frac{5 \times 20}{50} \right]$$

$$= \text{Rs.} [15 + 2] = \text{Rs.} 17$$

$$Z = \text{Rs.} 17$$

Here we see that the value of mean is less than the value of median ; value of median is less than the value of mode. That is, mean, median and mode are in ascending order of magnitude. Hence the given distribution is negatively skewed.

To calculate Pearson's coefficient of skewness we need standard deviation and we get it as follows;

Let us take A to be 17.5,

$$A = 17.5 \quad c = 5$$

Midvalue x	Frequency f	$d = \frac{x - A}{c}$	fd	\hat{d}^2	fd^2
2.5	10	-3	-30	9	90
7.5	40	-2	-80	4	160
12.5	30	-1	-30	1	30
17.5	90	0	0	0	0
22.5	20	1	20	1	20
27.5	10	2	20	4	40
Total	200		-100		340

Example 1:

$$N = \text{Total Frequency} = 200$$

$$\Sigma fd = -100$$

$$\Sigma fd^2 = 340$$

$$\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left[\frac{\Sigma fd}{N}\right]^2} \times c$$

$$= \sqrt{\frac{340}{200} - \left[\frac{-100}{200}\right]^2} \times 5$$

$$= \sqrt{\frac{17}{10} - \left[\frac{-1}{2}\right]^2} \times 5 = \sqrt{\frac{17}{10} - \left[\frac{1}{4}\right]} \times 5$$

$$= \sqrt{1.7 - .25} \times 5$$

$$= \sqrt{1.45} \times 5$$

Value of $\sqrt{1.45}$ is found out as follows :

$$\log 1.45 = .1614$$

$$\frac{\log 1.45}{2} = \frac{.1614}{2} = .0807$$

$$\text{Anti-log } (.0807) = 1.204$$

$$\sqrt{1.45} = 1.204$$

$$\therefore \sigma = 1.204 \times 5 = \text{Rs. } 6.02$$

$$\text{Standard deviation} = \text{Rs. } 6.02$$

$$\text{Karl Pearson's Coefficient of skewness} = \frac{3(a - M)}{\sigma}$$

$$= \frac{3(15 - (16.11))}{6.02} = \frac{3 \times (-1.11)}{6.02}$$

$$= \frac{-3.33}{6.02}$$

$$= -.553$$

Answer:

$$\text{Karl Pearson's coefficient of skewness} = -.553$$

Now we give below the second type of measure of skewness which is given in terms of quartiles.

6.5 Second type of measure of skewness :

The second measure of skewness is based on the fact that in a skewed distribution median does not lie half way between the lower and upper quartiles. That is, the second measure of skewness is based on the fact that the value of $(Q_3 - M) - (M - Q_1)$ is not equal to zero in a skewed distribution. Thus,

$$\begin{aligned}\text{Quartile measures of skewness} &= (Q_3 - M) - (M - Q_1) \\ &= (Q_3 - 2M + Q_1) \\ &= (Q_3 + Q_1 - 2M)\end{aligned}$$

The measure given above is an absolute measure of skewness.

6.6 Relative measure of skewness - Bowley's coefficient:

The relative measure is got by dividing this absolute measure by the sum $(Q_3 - M) + (M - Q_1)$

$$\begin{aligned}\therefore \text{Quartile coefficient of skewness} &= \frac{(Q_3 - M) - (M - Q_1)}{(Q_3 - M) + (M - Q_1)} \\ &= \frac{(Q_3 + Q_1 - 2M)}{(Q_3 - Q_1)}\end{aligned}$$

This Quartile coefficient of skewness is also known as Bowley's coefficient of skewness or Bowley's measure of skewness. This is a pure number.

6.7 Interpretation of the values of Bowley's coefficient :

For a symmetrical distribution the numerator viz., $(Q_3 + Q_1 - 2M)$ is zero and hence the coefficient of skewness is zero. It has positive or negative sign according as the given distribution is positively skewed or negatively skewed.

Value of the coefficient always lies between +1 and -1.

6.8 Drawback of the quartile measure and Coefficient of Skewness :

In some cases where the given distribution is not perfectly symmetrical, the value of $(Q_3 + Q_1 - 2M)$ may be zero. This is so because the values Q_3 and Q_1 are not based on the values of all the given items. Thus this measure of skewness should be used with great caution. For purposes of comparison, Karl Pearson's coefficient only should be used.

Example 1:

Space for hints

Consider the same problem given under Example-1 and calculate the quartile measure of skewness and its coefficient.

Wages in Rs.	Frequency
0 – 5	10
5 – 10	40
10 – 15	30
15 – 20	90
20 – 25	20
25 – 30	10

We have already calculated the value of median to be Rs. 16.11.

∴ M = Rs. 16.11.

Now we proceed to calculate the lower and upper quartiles are follows :

Class	Frequency	Cumulative Frequency
0 – 5	10	10
5 – 10	40	50
10 – 15	30	80
15 – 20	90	170
20 – 25	20	190
25 – 30	10	200

First we calculate the lower quartile as follows :

$$N = \text{Total frequency}$$

$$= 200$$

$$\frac{N}{4} = \frac{200}{4} = 50$$

$$Q_1 = \text{value of the item} \left[\frac{N}{4} \right]$$
$$= \text{value of the item (50)}$$

From the cumulative frequency column of the table given above, we come to know that all the items beyond the item 10 and upto the item 50 are having their values in interval "5–10".

The item 50 has its value in the interval "5-10" i.e., Q_1 has its value in the interval "5-10"

∴ "5-10" is the Q_1 class

l = true lower limit of the Q_1 class = 5

m = cumulative frequency of the class just above the Q_1 class = 10

f = Frequency of the Q_1 class = 40

c = magnitude of the Q_1 class = $10 - 5 = 5$

$$\therefore Q_1 = l + \frac{\frac{N}{4} - m}{f} \times c$$

$$= 5 + \frac{50 - 10}{40} \times 5$$

$$= 5 + \frac{40}{40} \times 5$$

$$= 5 + 5 = 10$$

Lower quartile = Rs. 10

Upper quartile is calculated as follows :

$$\frac{3N}{4} = \frac{3 \times 200}{4} = 150$$

$$Q_3 = \text{value of the item} \left[\frac{3N}{4} \right]$$

$$= \text{value of the item (150)}$$

From the cumulative frequency column of the above table we come to know that all the items beyond the item 80 and upto the item 170 are having their values lying in the interval "15-20".

The item 150 is in between the items 80 and 170

∴ Its value also lies in the interval "15-20".

i.e., value of Q_3 lies in the interval "15-20"

∴ "15-20" is the Q_3 class

l = true lower limit of the Q_3 class = 15

m = cumulative frequency of the class just above the

2.1 Percentile measure of Q_3 class = 80

f = frequency of the Q_3 class = 90

c = magnitude of the Q_3 class = $20 - 15 = 5$

$$Q_3 = l + \frac{\frac{3N}{4} - m}{f} \times c$$

$$= 15 + \frac{150 - 80}{90} \times 5$$

$$= 15 + \frac{70}{90} \times 5$$

$$= 15 + \frac{35}{9}$$

$$= 15 + 3.9 \text{ approx.}$$

$$= 18.9$$

∴ Upper quartile = Rs. 18.9

$$Q_1 = \text{Rs. } 10 \quad M = \text{Rs. } 16.11 \quad Q_3 = \text{Rs. } 18.9$$

Now we find out the quartile measure of skewness as follows :

$$\text{Quartile measure of skewness} = (Q_3 + Q_1 - 2M)$$

$$= \text{Rs. } [(18.9 + 10 - 2 \times 16.11)]$$

$$= \text{Rs. } (28.9 - 32.22)$$

$$= \text{Rs. } (3.32)$$

Quartile coefficient of skewness is found out as follows :

$$\text{Quartile coefficient of skewness} = \frac{(Q_3 + Q_1 - 2M)}{(Q_3 - Q_1)}$$

$$= \frac{18.9 + 10 - (2 \times 16.11)}{18.9 - 10}$$

$$= \frac{28.9 - 32.22}{8.9} = \frac{-3.32}{8.9}$$

$$= -0.37$$

Space for hints

Answer:

Quartile measure of skewness = Rs. (3.32)

Quartile coefficient of skewness = $-.37$

7. Application of Skewness in Economics

Many frequency distributions derived from economic data are extremely skewed.

Measures of skewness are widely used to study the inequalities in the distribution of income, wealth, wage etc. Wherever human activities are concerned the distribution is mostly skewed.

The American income distribution is positively skewed. The male-female differences in income levels and occupational differences are highly responsible for this skewness.

Indian income distribution also exhibits a high degree of positive skewness. Landholding pattern, nature of distribution of urban property holdings are the main causes for such a skewed distribution of income.

The distribution of firms within an industry according to the number of employees or total sales is found to be skewed.

Wealth is another important economic variable that shows skewness.

In biological studies and other studies depending more or less upon the laboratory experiments, distributions mostly tend to be symmetrical. But in social and economic enquiries a perfectly symmetrical distribution is an exception, and a large degree of skewness is generally present.

III. Measures of Kurtosis

1. Kurtosis - Meaning

The flatness or peakedness of a frequency curve is known as Kurtosis. Flatness or peakedness of a curve depends upon the number of values of items near the value of mode. We judge the flatness or peakedness of a curve with respect to the normal curve. Normal curve is an ideal symmetrical curve having many more mathematical properties. Measures of Kurtosis tell us how near a particular frequency curve conforms to the normal curve.

2. Measure of Kurtosis

There are various measures to measure the Kurtosis of a frequency distribution. Of these measures, the percentile measure of Kurtosis is the simplest one. All the other measures are more mathematical.

2.1 Percentile measure of Kurtosis

Space for hints

The percentile measure of Kurtosis is obtained by dividing the difference between the 10th and 90th percentiles by the quartile deviation. Therefore,

$$\begin{aligned}\text{Percentile measure of Kurtosis} &= \frac{\text{Difference between 10th and 90th percentiles}}{\text{Quartile deviation}} \\ &= \frac{(P_{90} - P_{10})}{\frac{(Q_3 - Q_1)}{2}}\end{aligned}$$

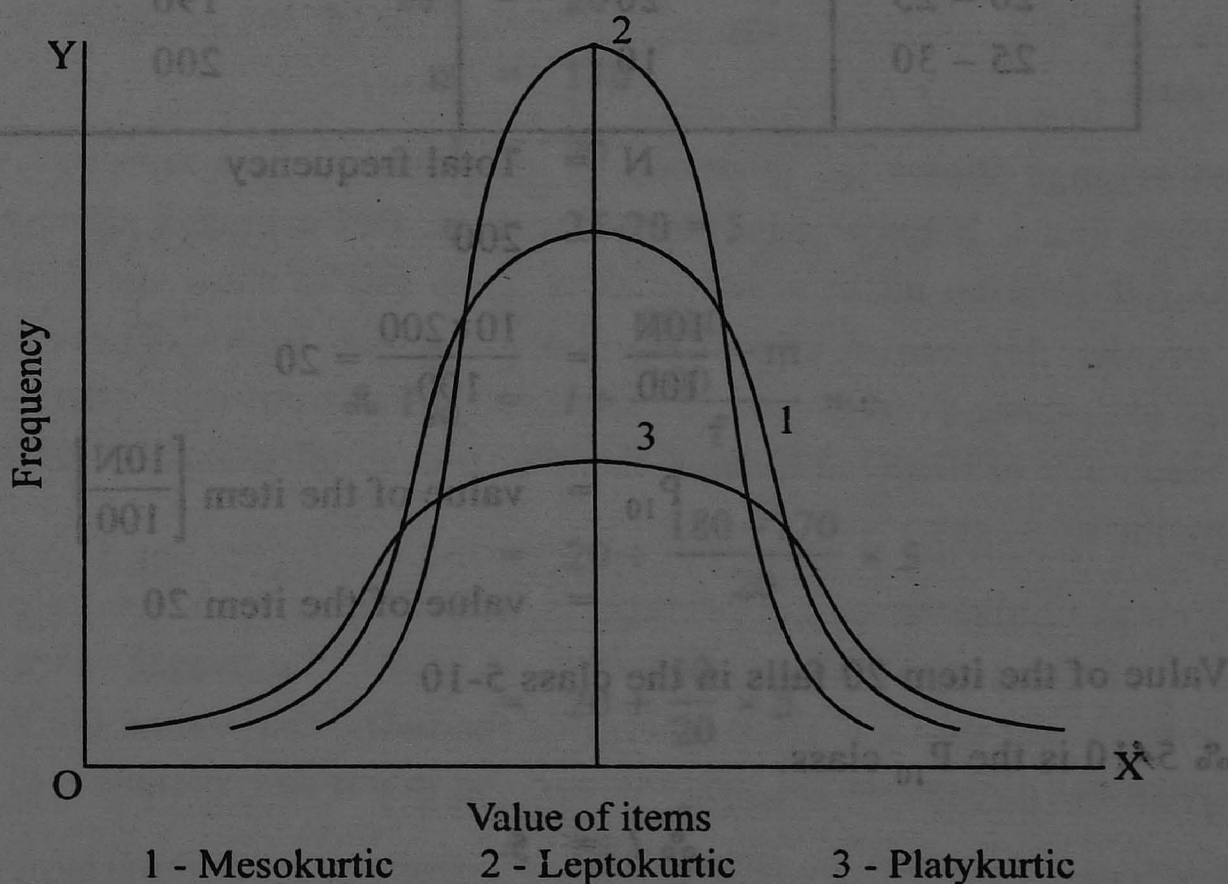
2.2 Interpretation of the values of Percentile measure - Types of Kurtosis

For the normal curve the value of percentile measure of Kurtosis is equal to 3.8. The normal curve, and any distribution for which the percentile measure of Kurtosis is equal to 3.8 are called "mesokurtic".

If the percentile measure of Kurtosis is greater than 3.8 for a given distribution, then the curve is said to be "leptokurtic". The graph of a leptokurtic curve will be more peaked than normal curve.

If the percentile measure of Kurtosis is less than 3.8 for a given distribution, then the curve is said to be "platykurtic". The graph of a platykurtic curve is more flat than the normal curve.

In the figure given below, the curve numbered as 1 is normal curve and it is mesokurtic. The curve numbered as 2 is leptokurtic and the curve numbered as 3 is platykurtic.



Example :

Calculate the measure of Kurtosis from the following frequency distribution:

Wages in Rs.	Frequency
0 – 5	10
5 – 10	40
10 – 15	30
15 – 20	90
20 – 25	20
25 – 30	10

In the previous topic in example-2 we have got the values of the two quartiles Q_1 and Q_3 .

$$Q_1 = \text{Rs. } 10 \quad Q_3 = \text{Rs. } 18.9$$

So, here we calculate P_{10} and P_{90} only and get the value of percentile measure of Kurtosis.

Class	Frequency	Cumulative frequency
0 – 5	10	10
5 – 10	40	50
10 – 15	30	80
15 – 20	90	170
20 – 25	20	190
25 – 30	10	200

$$N = \text{Total frequency}$$

$$= 200$$

$$\frac{10N}{100} = \frac{10 \times 200}{100} = 20$$

$$P_{10} = \text{value of the item } \left[\frac{10N}{100} \right]$$

$$= \text{value of the item } 20$$

Value of the item 20 falls in the class 5-10

∴ 5-10 is the P_{10} class.

$$\therefore l = 5$$

$$m = 10$$

$$f = 40$$

$$c = 10 - 5 = 5$$

$$\therefore P_{10} = l + \frac{\left[\frac{10N}{100} - m \right]}{f} \times c$$

$$= 5 + \frac{(20 - 10)}{40} \times 5$$

$$= 5 + \frac{10}{40} \times 5$$

$$= 5 + \frac{5}{4}$$

$$= 5 + 1.25 = 6.25$$

$$\frac{90N}{100} = \frac{90 \times 200}{100} = 180$$

$$P_{90} = \text{value of the item } \frac{90N}{100}$$

$$= \text{value of the item } 180$$

\therefore 20-25 is the P_{90} class

$$\therefore l = 20$$

$$m = 170$$

$$f = 20$$

$$c = 25 - 20 = 5$$

$$\therefore P_{90} = l + \frac{\frac{90N}{100} - m}{f} \times c$$

$$= 20 + \frac{180 - 170}{20} \times 5$$

$$= 20 + \frac{10}{20} \times 5$$

$$= 20 + \frac{5}{2}$$

$$= 20 + 2.5$$

$$= 22.5$$

$$\begin{aligned} \text{Percentile measure of Kurtosis} &= \frac{(r_{90} - P_{10})}{(Q_3 - Q_1)} \\ &= \frac{22.5 - 6.25}{18.9 - 10} \\ &= \frac{16.25}{8.9} \times 2 \\ &= \frac{32.50}{8.9} \\ &= 3.6 \text{ approx.} \end{aligned}$$

Percentile measure of Kurtosis for the given distribution is less than 3.8. Therefore, the curve of the given distribution is platykurtic.

IV. Averages, Measures of Dispersion, Skewness and Kurtosis - Contrasted

One of the purposes of statistics is that of describing distribution in precise mathematical terms in order to study various properties of a given distribution. The measures of central tendency, dispersion, skewness and kurtosis are the various sets of measures used to describe a given distribution.

An average shows the tendency of the group. It is a representative of a distribution and it helps to summarize the data. But averages alone will not give full information about a distribution. They fail to show the form of the series or the degree of variability between the items. There may be distributions whose averages are the same but differ from each other in many ways. Measures of dispersion show this and this supplement the information given by the averages.

Typical character of an average is determined with the help of the measures of dispersion. When dispersion is small, the average is a typical value in the sense that it closely represents the individual value. On the other hand, when the dispersion is large, the average is not so typical.

Averages serve as a useful tool for purposes of comparison of two

distributions. Thus they give a mathematical relationship between the different distributions. Here the measures of dispersion act as a useful check on drawing wrong conclusions from the comparison of the averages. Also, if two distributions are in two different units, comparison is not possible using averages. But with the help relative measures of dispersion, all such comparisons can be easily made.

Measures of dispersion may also be used to estimate the value of a series itself.

Two different distributions may have the same averages and measures of dispersion but the distributions may still differ. Also, the scatter of data on either side of the measures of central tendency may not be same for both. The scatter of data on either side of an average may be symmetrical or not. Measures of dispersion give only the extent of variation and not about the symmetry. The extent of asymmetry is given by the measures of skewness.

Thus the measures of dispersion study the size of the distribution, while the skewness studies the shape of the distribution.

11. Answers to the Check Your Progress Questions :

- | | |
|---------------|---------------------|
| 1. Refer 1 | 12. Refer 7.5 |
| 2. Refer 3 | 13. Refer 8.1 |
| 3. Refer 4 | 14. Refer 8.3 |
| 4. Refer 5.1 | 15. Refer 8.4 |
| 5. Refer 5.2 | 16. Refer 8.5 |
| 6. Refer 6.1 | 17. Refer 8.9 |
| 7. Refer 6.2 | 18. Refer 8.10 |
| 8. Refer 7.1 | 19. Refer 8.11 II D |
| 9. Refer 7.2 | |
| 10. Refer 7.3 | |
| 11. Refer 7.4 | |

12. Model questions for guidance :**10 Marks Questions (One Page Answer)**

1. Write short notes on Quartile Deviation.
2. Marks obtained by ten students are given below.
50, 55, 57, 49, 54, 61, 64, 59, 59, 56
Calculate the quartile deviation.
3. The price in Rupees of a certain commodity quoted in a market on the different months of calendar year is given below :
25, 35, 55, 63, 27, 45, 57, 23, 50, 61, 38
Calculate the range and the coefficient of range.
4. From the following data compute mean deviation :

Size	6	9	12	15	18
Frequency	14	24	38	20	4

5. Calculate (1) S.D. (2) C.V. in respect of marks obtained by 10 students given below :
50, 55, 57, 49, 54, 61, 64, 59, 58, 56.
6. Given the following figures of production (in tons) during consecutive weeks in a chemical factory.
120, 142, 103, 138, 126, 113, 132, 122
Obtain C.V.
7. What are the measures of dispersion? Why is S.D. taken as the best measure?
8. Critically examine the claims of the M.D. and S.D. as rival measures of dispersion. Why are they not calculated using the same average?
9. What do you mean by dispersion? How is it different from an average?
10. What is skewness? What are its types? Explain.
11. What is Kurtosis? How do you measure it?

20 Marks Questions (Three Page Answer)

Space for hints

1. What is measure of dispersion ? What are its purposes ? Explain the necessity of measures of dispersion when there are averages.

2. Calculate semi-inter quartile range and quartile coefficient of dispersion from the following data.

Age in years	20	30	40	50	60	70	80
No. of members	3	61	132	153	140	51	3

3. The following table gives the distribution of monthly income of 600 Middle class families in a city.

Monthly income Rs.	No. of families
0 – 75	69
75 – 150	167
150 – 225	207
225 – 300	65
300 – 375	58
375 – 450	54
450 – 525	10

Calculate the quartile deviation.

4. Calculate the quartile deviation of wages.

Wages in Rs.	30-32	32-34	34-36	36-38	38-40	40-42	42-44
Labourers	12	18	16	14	12	8	6

5. Marks obtained by ten students are given below:

50, 55, 57, 49, 54, 61, 64, 59, 58, 56.

Calculate i) M.D. from mean ii) M.D. from median.

6. Calculate M.D from mean and median for the following data :

100, 150, 200, 250, 360, 490, 500, 600, 671.

Also calculate coefficient of mean deviation.

7. Age distribution of hundred life insurance policy holders is as follows :

Age	17-19	20-25	26-35	36-40	41-50	51-55	56-60	61-70
Number	9	16	12	26	14	12	6	5

Calculate mean deviation from the median age and coefficient of mean deviation.

8. Compute the standard deviation from the following series of electricity consumption

Consumption in K.W.	0-10	10-20	20-30	30-40	40-50	50-60
No. of users	10	25	65	85	15	100

9. Compute the standard deviation from the following data giving the age distribution of the parliament members.

Age in years	20-30	30-40	40-50	50-60	60-70	70-80	80-90
Members	2	61	132	153	140	51	2

10. Calculate the coefficient of variation for the following distribution of wages per week in a factory.

Wages per week in Rs.(midvalue)	38	44	50	56	62	68	74	80	86
Frequency	27	72	135	170	285	175	96	28	12

11. An analysis of the monthly wages gives the followed results.

	Firm A	Firm B
No. of workers	500	500
Average monthly wage	186	175
Variance of the distribution of wages	81	100

12. Frequency distribution of monthly earnings of workers in a factory is as follows :

Rs.	80.00	92.40	83.20	91.80	84.30	85.00	93.70
Frequency :	7	7	45	43	28	14	46

Find the mean and S.D. of their earnings.

13. Using coefficient of variation, state which of the series is more variable.

Space for hints

A	125	134	131	128	127	127	129	120	130	131
B	120	118	119	118	112	117	121	118	121	120

14. Which is the best measure of dispersion? Why is it called like that?

Ans : Hints.

- (i) First give a general account of dispersion.
- (ii) Characteristic of a good measure of dispersion.
- (iii) Various measures of dispersion.
- (iv) Then explain with examples, how S.D. satisfies most of the characteristics explained above in (ii), while other measures do not satisfy most of them. Thus show that the S.D. scores over all the other measures of dispersion and it is the best measure.

15. Write short notes on :

- (i) Skewness
- (ii) Kurtosis

16. Calculate Pearson's coefficient of Skewness.

x	y
10-20	12
20-30	18
30-40	35
40-50	42
50-60	50
60-70	45
70-80	20
80-90	8

